

# The role of cognitive factors on the development and evolution of the vocabulary

James Brand

Thesis submitted for the degree of Doctor of Philosophy

Department of Psychology

Lancaster University

March 2017





## Abstract

This thesis aims to explore how psycholinguistic factors can play a pivotal role in the way that the human lexicon is structured. Recent approaches to studying language have been heavily influenced by the principle that languages are shaped to suit the general learning and processing biases of the learner ([Christiansen & Chater, 2008](#)). This view has gained considerable support from theoretical and empirical accounts from researchers in the fields of language acquisition and evolution. However, there has been little evidence that demonstrates how such learning biases may operate differently at various stages of an individual's vocabulary development, or indeed how they may operate differently as the vocabulary subsequently evolves. Through a series of experimental studies, this thesis will examine how well documented psycholinguistic properties of words can shape the lexicon throughout the course of a language learner's life, as well as the life of the language itself.

Chapters 2 and 3 investigate how both arbitrariness and non-arbitrariness (in the form of sound-symbolism) make significant contributions to the way that languages are learnt. Central to this point is that important considerations should be made as to when and how these two properties are considered beneficial. Principally, these chapters focus on how the size of the vocabulary directly influences whether arbitrariness or non-arbitrariness is advantageous for the learner. We show that sound-symbolism facilitates learning individual form-meaning mappings when the vocabulary size is small, whereas as the vocabulary grows this advantage is only observed for distinguishing between categories. Which may explain why words acquired earlier in life, when the vocabulary size is small, are reported to be more sound-symbolic ([Monaghan et al, 2014](#)).

Chapter 4 goes on to examine how variation in psycholinguistic properties of word can be used to directly predict variation in the fidelity of the word's learning and production. By manipulating the frequency, length and age of acquisition of words learnt within an artificial language learning paradigm, this chapter aims to show how certain words are less vulnerable to change based on the way these properties offer significant processing advantages, whilst also considering the nature of the way errors are produced during recall by distinguishing between lexical adjustment and replacement. Importantly, these properties are investigated independent of one another, allowing for a significant contribution to be made to the way language change is studied.

Following on from this, Chapter 5 then considers how these psycholinguistic properties may come to change the lexicon over the course of evolution. By using an

iterated learning paradigm, the aim here is to observe how the changes reported in Chapter 4 may explain previously reported differences in the stability, and indeed instability, of lexical forms over a much longer timescale. This approach examines such differences through the cultural transmission of the languages across several generations of learners..



## Acknowledgements

Whilst this thesis has been the culmination of much hard work, it would not have been possible without the help and support of many people. Firstly, my supervisors Prof. Padraic Monaghan and Dr. Peter Walker, who have always been exceptionally supportive and willing to share their time and thoughts with me. I would like to highlight Padraic's role in the progression and development of this thesis in particular, his astonishingly vast knowledge has helped guide much of this work, constantly demonstrating an indefatigable passion for research (and tea), which I can only hope to emulate. I have been fortunate to be part of his lab group and more broadly part of LuCiD, where many great discussions and insights have been generated by an exceptionally bright group of individuals. Most notably Rebecca Frost, who I have had the great privilege to call a friend as well as a colleague, from day one she has always offered her help, advice and motivation, undoubtedly keeping my spirits high throughout the PhD, thanks science friend!

Whilst my academic career is only in its early stages still, there have been a number of people who have played a critical role in the journey that got me to this point. My high school English teacher, Miss Sweet, who provoked a real interest in words and how fascinating they can be. Dr. Peter Jones, my Undergraduate supervisor, who introduced me to the science of communication and an all-round great guy. Also Dr. Chrissy Cuskley, who introduced me to the fascinating words *bouba* and *kiki*, her support during my Masters has been pivotal in why I chose to undertake a PhD. Similarly, there has been much help from technology too, without publically available resources such as Stackoverflow, Statmethods, Wikipedia and even the humble Google Scholar, much of this work would have been considerably more difficult.

I have thoroughly enjoyed my time at Lancaster, making some remarkably good friends in the process, the overall environment and camaraderie here has been fantastic, I have always felt at home and part of the department, which has made the PhD all the more enjoyable. Whilst there are too many people to name, I feel it is important to acknowledge a few, in particular Christian, Helen, Katie, Kay, Lara, Liam, Malen, Matt, Steven N, Steven W and Han (who has put up with me the most), all of whom have been exceptional friends, for which I am truly grateful. I would also like to thank Louise Brand, a brilliant mother who has always been supportive of my endeavours and has given me endless encouragement, morale boosts and discussions about the weather. Also, to my dear friends of 297 Abbeydale Road, who have always been there to share the music of Mark Morrison with me. Finally, my OCD, whilst it has made aspects of this work difficult to complete, facing up to the challenges it presents and overcoming them is a hugely rewarding feeling

I'd like to dedicate this thesis to Adrian & Alan Neville.

### **Declaration**

I hereby declare that this thesis is my own work, and has not been submitted in substantially the same form for the award of a higher degree at this institution or elsewhere.

This research was supported by funding from the Department of Psychology, Lancaster University from 1st October 2013 until 30th September 2016.

I also declare that parts of this thesis have been submitted to academic journals for publication. These publications are as follows:

Brand, J., Monaghan, P. & Walker, P. (Under review). The Changing Role of Sound-Symbolism for Small Versus Large Vocabularies. *Cognition*.

James Brand

# Contents

---

<b>Abstract</b>	<b>i</b>
<b>Acknowledgments</b>	<b>iii</b>
<b>List of Tables</b>	<b>vii</b>
<b>List of Figures</b>	<b>viii</b>
<b>1 Cognitive Factors Shaping the Vocabulary</b>	<b>1</b>
1.1 Thesis Overview	1
1.2 A Taxonomy for Arbitrariness and Non-arbitrariness in Natural Language .....	5
1.2.1 Arbitrariness .....	6
1.2.2 Iconicity .....	7
1.2.3 Systematicity .....	13
1.3 Division of Labour in the Developing Vocabulary .....	16
1.3.1 Learnability .....	16
1.3.2 Usability .....	21
1.4 Cognitive Perspectives on Lexical Evolution and Change .....	25
1.5 Thesis Outline .....	30
<b>2 The Changing Role of Sound-Symbolism for the Growing Vocabulary</b>	<b>32</b>
2.1 Introduction .....	32
2.2 Experiment 1 – Learning From a Congruent and Incongruent Language .....	37
2.3 Discussion .....	43
<b>3 Form-meaning Mappings and the Structure of the Vocabulary</b>	<b>47</b>
3.1 Introduction .....	47
3.2 Experiment 2 – Fully Sound-symbolic Language .....	55

3.3 Experiment 3 – Fully Arbitrary Language .....	62
3.4 Across Experiments Comparison .....	69
3.5 General Discussion .....	72
<b>4 Predictors of lexical stability in an artificial language</b>	<b>74</b>
4.1 Introduction .....	74
4.2 Experiment 4 – Frequency Effects .....	77
4.3 Experiment 5 – Word Length Effects .....	96
4.4 Experiment 6 – Age of Acquisition Effects .....	109
4.5 General Discussion .....	123
<b>5 Predictors of lexical stability through the cultural transmission of language</b>	<b>126</b>
5.1 Introduction .....	126
5.2 Experiment 7 – Frequency Effects and Language Evolution .....	131
5.3 Experiment 8 – Word Length Effects and Language Evolution .....	144
5.4 Experiment 9 – Age of Acquisition Effects and Language Evolution .....	153
5.5 General Discussion .....	160
<b>6 Conclusions</b> .....	<b>163</b>
6.1 Aims of Thesis .....	163
6.2 Summary of Findings .....	165
6.3 Limitations and Future Directions .....	172
6.4 Concluding Remarks .....	177
<b>References</b>	<b>179</b>
<b>Appendix</b>	<b>207</b>

# List of Tables

---

2.1	List of phonetically transcribed words used during experiment 1	38
3.1	List of phonetically transcribed non-words used during experiment 3	65
4.1	Example language sets used in Experiments 4, 5 and 6	86
4.2	Contingency table showing counts of different response types in Experiment 4	92
4.3	Contingency table showing counts of different response types in Experiment 5	106
4.4	Training and testing procedure for Experiment 6	117
5.1	Example of languages produced in Experiment 7	138
A1.1	Main model selection for Experiment 1	207
A1.2	Summary of the Generalized Linear Mixed-effects Model of (log odds) accuracy of responses for Experiment 1	208

# List of Figures

---

1.1	Examples of different types of non-arbitrary form-meaning mappings	8
1.2	Phonological cues reported to distinguish between open and closed class words from Monaghan et al. (2011)	15
1.3	Iconicity, systematicity and arbitrariness in a small vocabulary size	19
1.4	Iconicity, systematicity and arbitrariness in a large vocabulary size	23
2.1	Examples of a same and different category trial from Experiment 1	36
2.2	Example of Likert scale item for correspondence between word and rounded or angular shapes	39
2.3	Mean accuracy of responses by block for Experiment 1	42
2.4	Mean accuracy of responses in categorical trials for Experiment 1	44
3.1	Mean accuracy of responses by block for Experiment 2	60
3.2	Mean accuracy of responses for Experiment 2	60
3.3	Mean accuracy of responses from questionnaire analysis of sound-symbolism	66
3.4	Mean accuracy of responses by block for Experiment 3	68
3.5	Mean accuracy of responses for Experiment 3	68
3.6	Mean accuracy of responses for fully congruent and no relationship experimental conditions	71
4.1	Results from Experiment 4	89
4.2	Calculation of critical threshold for adjustment/replacement classification	91
4.3	Types of change for Experiment 4	93
4.4	Results from Experiment 5	104
4.5	Types of change for Experiment 5	107
4.6	Results from Experiment 6	119

4.7	Types of change for Experiment 6	120
5.1	Visual representation of the iterated learning model taken from Roberts (2013)	132
5.2	Mean error of responses in Experiment 7	137
5.3	Types of change for Experiment 7	141
5.4	Analysis of word length changes in Experiment 8	146
5.5	Mean error of responses in Experiment 8	148
5.6	Types of change for Experiment 8	150
5.7	Mean error of responses in Experiment 9	155
5.8	Types of change for Experiment 9	157
A2.1	Mean accuracy of responses for Experiment 1 & 2	211
A2.2	Mean accuracy of responses for Experiment 1 & 3	214





# Chapter 1

## Cognitive factors shaping the vocabulary

---

### 1.1 Thesis overview

Words are a fundamental unit of human language, enabling the representation of specific meanings through the production of a discrete signal. Across the world users of a language will make use of a vocabulary comprising thousands of words, with each word form highly likely to differ across languages. But what determines why a particular word form should come to represent a specific meaning? Indeed, such a question has been debated throughout history, with [Plato's \(1971\)](#) *Cratylus* prompting contemplation on the topic, raising the question as to whether names are themselves meaningful or simply used to signify their meaning. More contemporary schools of thought have placed much stronger emphasis on the idea that words exhibit, or even demand, an arbitrary relationship between their form and meaning ([de Saussure, 1916](#); [Hockett, 1960](#)).

Currently however, there is an ever growing body of research that is shedding new light on the extent to which non-arbitrary mappings are present in

language, advancing our understanding of the potential benefits such mappings can offer when a language must undergo acquisition, processing, change and evolution. Take for example [Harnad's \(1990\)](#) Symbol Grounding Problem. How could a communicative system such as language arise, when it uses seemingly arbitrary forms to represent meanings, when the forms themselves must be defined by more arbitrary forms (see also, [Oliphant, 2002](#)). A non-arbitrary system however, could offer a neat solution to such a problem, whereby the form itself provides information about the meaning, with recent research uncovering substantial evidence of the existence of such non-arbitrariness in natural language.

Often referred to as sound-symbolism, it has been proposed that language users can exploit the dissociation between form and meaning to ground their communicative system ([Imai & Kita, 2014](#); [Ramachandran & Hubbard, 2001](#)). Consider a scenario where somebody is trying to communicate the meaning of 'dog' to somebody who does not share the same language. The phonological sounds comprising the English word form *dog* offers very little information about the intended referent. In contrast, using a sound-symbolic form, such as *woof*, where the sound of the dog itself is imitated, carries within the form itself referential information about what the intended meaning is.

However, one must also acknowledge the potential advantages that arbitrariness offers to the vocabulary if a fully coherent account of that vocabulary is to be presented. If non-arbitrariness dominates the vocabulary, then it potentially introduces problems for the language user. For instance,

expressivity could be restricted when relying upon a system that uses one-to-one mappings between form and meaning, whereby finding a suitable form to represent more abstract or complicated meanings would prove difficult (e.g. it would be hard to find a way to reliably represent the meaning *perhaps* with a non-arbitrary form, but with arbitrariness it is relatively simple). In addition, a vocabulary that can communicate thousands of different meanings has to ensure that ambiguity does not cause unnecessary ambiguity, arbitrariness provides a neat solution to this problem as similar meanings do not require similar forms. However, a non-arbitrary system would be dependent on similar forms for similar meanings, potentially creating an inefficient system for the user, where words would be difficult to disambiguate between (i.e. finding non-arbitrary forms to represent the many different types of bird in the world would require a complex solution to resolve any ambiguity) (Gasser, 2004; Monaghan, Christiansen & Fitneva, 2011).

Thus, understanding the dynamics of a vocabulary and the influence of these two options for mapping form to meaning is of fundamental importance. A primary aim of this thesis will be to present a view of the vocabulary which considers the dynamic requirements of a language user, whereby the advantages of arbitrariness and non-arbitrariness do not compete against each other in parallel, but instead exert influence at different stages of development and evolution. This will be addressed in Chapters 2 and 3 through a series of experimental studies, which examine the role of sound-symbolism and

arbitrariness in language learning, when the dynamics of a growing vocabulary are introduced.

Another core aim of the thesis will be to explore a recent issue of contention in the literature surrounding the link between the acquisition and the evolution of the vocabulary ([Christiansen & Chater, 2008](#); [Croft, 2000](#)). If the presence of sound-symbolism does in fact contribute significantly to the bootstrapping of communication systems (as proposed by Imai and Kita (2014)), in addition to their relative processing benefits, then such forms should be afforded a privileged status within the vocabulary. This would mean that the acquisition process directly influences the evolution of the vocabulary, with words acquired earlier in life, not only incorporating sound-symbolic features, but also ensuring that such features are conserved in the language. In contrast to those words acquired later on in life, which incorporate substantially less sound-symbolism ([Massaro & Perlman, 2017](#); [Monaghan, Shillcock, Christiansen & Kirby, 2014](#), [Perry, Perlman & Lupyan, 2015](#)), which would exhibit less resistance to substantial changes in their phonology over time. This will be addressed through a series of experiments in [Chapters 4 and 5](#), which examine the ways that psycholinguistic properties of words (such as age of acquisition, frequency and word length) can influence the way those words are processed, and crucially that they also can be used to predict variation in the rate of lexical evolution.

In this introductory chapter, I will outline the core theoretical arguments and empirical evidence motivating the central aims of the thesis. I will begin in

section 1.2 by establishing a taxonomy of arbitrariness and non-arbitrariness, in order to present and conceptualise a highly textured view of the vocabulary. In section 1.3, the benefits and limitations of both arbitrary and non-arbitrary systems for a vocabulary will be discussed, with a focus on the pressures introduced by the language learner and also the language user. Developing on from that discussion, section 1.4 will focus on presenting a cognitive explanation for processes of lexical evolution and change, particular attention will be given to the way that acquisition can directly influence evolution.

## **1.2 A taxonomy for arbitrariness and non-arbitrariness in natural language**

Within human language there are different relations which map forms onto meanings, these being arbitrariness and non-arbitrariness (iconicity and systematicity). Historically, there has been weighted emphasis placed on the arbitrariness of this mapping ([de Saussure, 1916](#)), with instances of non-arbitrariness considered however there has been a shift in this position more recently. In current research, understanding the complexities of how both arbitrariness and non-arbitrariness can manifest themselves within human language, has been of considerable importance. Thus, viewing the vocabulary as a system that is not exclusively arbitrary (or even exclusively non-arbitrary), has enabled advances in our understanding of the way the vocabulary is structured. It is therefore imperative that these two relationships, and the distinct forms they occur in, are clearly defined and conceptualized.

### 1.2.1 Arbitrariness

When we produce a word, the word's form will likely not hold any information about the intended meaning. Such a relationship can be considered arbitrary – it is only through the conventionalised use of language that allows users to understand what the word refers to. Likewise, an arbitrary relationship between form and meaning also means that word forms could easily be interchanged, as long as the language users have a shared comprehension of what the form is referring to. Take for instance Hockett's (1960) neat example of the word *whale*, it is represented by the phonological form /weɪl/ in English, yet this has no clear relationship to its meaning, i.e. a large marine mammal. In many different languages the word *whale* has come to be represented through a variety of different phonological forms, such as in Finnish /valas/, Spanish /baʎena / or /kít/ in Slovene. Indeed, arbitrariness means that any of these forms can be learnt and then used, as long as the form has been adopted by users of the language.

Another hallmark of arbitrariness in human language is the unstructured relationship between forms that share similar meanings. If we consider the word forms *dog* and *coyote*, both are used in English to refer to animals classed under the genus *Canis*, share many physical and behavioural characteristics, yet are separate species. Despite the two animals being highly related by their shared semantic information, their phonological representations are noticeably distinct from one another. Similarly, if we compare the word forms *dog* and *frog*, the two phonological forms are now substantially more similar, yet each of their respected meanings are considerably more distinct from each other. This can also

be extended to homophones, such as *new* and *knew*, where an identical phonological form can be used to express two individual meanings. Thus, within an arbitrary relationship between form and meaning, there is no systematic relationship which ties phonological features to semantic meaning.

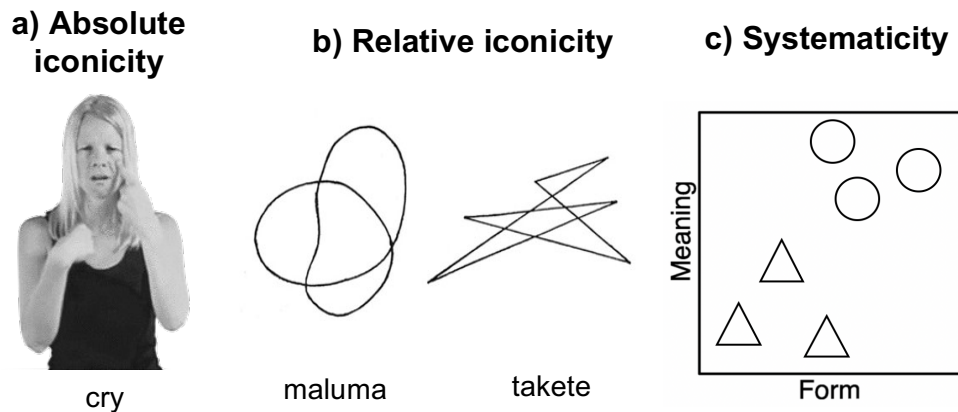
Arbitrariness in language has long been considered a fundamental feature of language (Hockett, 1960), which relies on the cognitive capabilities of humans to successfully learn and use symbolic forms in communication, something that has been proposed to be a human unique behaviour (Deacon, 1997). However, the supposed exclusivity for arbitrariness in language has recently been challenged, with substantial evidence revealing a much more prominent role and potential for non-arbitrary mappings being incorporated within human language (Dingemanse, Blasi, Lupyan, Christiansen & Monaghan, 2015).

### 1.2.2 Iconicity

Perhaps the most well documented form of non-arbitrariness in language comes in the form of iconic mappings. Although iconicity has been demonstrated in many different ways, the defining feature of an iconic mapping is the way the form provides a resemblance to its intended meaning, normally through perceptuomotor analogies (Dingemanse et al, 2015; Schmidtke, Conrad & Jacobs, 2014). There are generally two main classifications of iconicity in the literature - absolute and relative<sup>1</sup> – each of which will be discussed in turn.

---

<sup>1</sup> Here, we follow the terminology adopted by Dingemanse et al (2015), which has been adapted from Gasser, Sethuraman and Hockema's (2010) categorisation. The terms used here are most relevant to the main themes of this thesis, however it should be noted that



**Figure 1.1.** Examples of different types of non-arbitrary form-meaning mappings. **a) *Absolute iconicity*** in British Sign Language for the word *cry* (adapted from [Perniss et al, 2010](#)) **b) *Relative iconicity*** using shape-sound associations (adapted from [Köhler, 1929](#)) **c) *Systematicity*** where word forms incorporate a cue that is used differentially across meaning categories, here triangles may represent nouns and circles represent verbs.

*Absolute iconicity* offers a direct link between form and meaning, with common examples being found within onomatopoeia. Here, iconicity often imitates the perceptual properties of a meaning, such as ‘*moo*’ being used to represent the sound that a cow makes. This type of iconicity can also be observed within non-verbal forms of communication, such as in sign language and communicative gestures, where visual features can be imitated directly ([McNeill, 1992](#); [Perniss, Thompson & Vigliocco, 2010](#)) or even through graphical representations ([Garrod, Fay, Lee, Oberlander & MacLeod, 2007](#)), see [Figure 1.1a](#). Indeed, absolute iconicity is constrained to a large extent by the sensory modality of the meaning being imitated, one would encounter difficulties producing an absolute

---

there are many different classifications used in the literature, for a more general discussion on these see [Lockwood and Dingemanse \(2015\)](#).



iconic form for the visual properties of a meaning such as *tree* in the auditory modality, whereas in the visual modality the task would be considerably easier.

Although the form used in an absolute iconic mapping has been suggested to demonstrate a one-to-one resemblance (for instance by [Dingemanse et al, 2015](#)), there is the possibility that the meaning being represented covers a much broader semantic classification. For instance, the onomatopoeic form *woof woof*, does indeed imitate the sound that a dog makes, however the term can also be used to refer to the dog itself. Therefore, it is important to note that with an absolute iconic form, there is the potential for metonymical extension of the meaning, whereby a much more general meaning is expressed ([Akita, 2013](#)). Additionally, within the vocabulary one is likely to find a more frequently used, conventionalized and arbitrary alternative, which can be used to express the same broad meaning as the absolute iconic form, in the case of the form *woof woof*, there is the word *dog* ([Laing, 2014](#)).

Indeed, one could even argue that although imitative, there is considerable diversity in the way absolute iconic word forms are produced across languages. The English form *woof woof* is used to represent the sound a dog makes, yet in Czech the form used is *haf haf* and in Japanese it is *wan wan*, demonstrating that such forms do not need to be universally identical. [Assaneo, Nichols and Trevisan \(2011\)](#) have claimed that phonological constraints, introduced by the anatomy, physiology and phonetic space of the human vocal system, have shaped the differences in onomatopoeic forms. This further highlights the fact that absolute iconicity is only an imitative form, not a perfect reduplication of a

naturally occurring perceptual property, and like arbitrary forms, are subject to conventionalization and variance across languages (Edmiston, Perlman & Lupyan, 2016; Perlman, Dale & Lupyan, 2015).

*Relative iconicity* is the other type of iconic mapping found in language. Here, the form provides an analogous relationship to the intended meaning. Mappings of this kind are much more varied than those that exhibit absolute iconicity, but as a general distinction between the two, relative iconicity tends to depict meaning through aspects of the form, whereas absolute iconicity attempts to imitate the meaning through the form. This enables relative iconicity to map forms onto more abstract and varied meanings.

One well researched example of relative iconicity comes in the form of ideophones, words that provoke vivid sensory imagery (Dingemanse, 2012). For example, the words *goro* and *koro* in Japanese are used to represent the meanings of a heavy and light object rolling respectively, with duplication also being used to extend the analogy to multiple heavy/light objects rolling (i.e. *gorogoro*). Ideophones are widely reported in many languages of the world, although they are notably less prevalent in many Indo-European languages. Yet, examples of mappings that are characteristic of relative iconicity, which appear to be cross-linguistically universal, have also been studied.

Such examples are usually referred to as cross-modal associations, where a perceptually salient feature of a meaning corresponds reliably to a feature in another modality. A well-documented example of this comes from Köhler's (1929) *takete/maluma* experiment, where these two non-words are reliably paired

by participants to angular/rounded shapes (*takete* with the angular shape), see [Figure 1.1b](#). Although a large variety of cross-modal associations have been reported (see [Spence, 2011](#) for review), much of the research only has indirect implications for our understanding of their use in natural language. However, there is evidence to suggest that cross-modal associations do influence the structure of the vocabulary, with [Monaghan, Mattock and Walker \(2012\)](#) demonstrating that rounded\angular distinctions are marked by differential uses of velar and voicing phonological properties in English. [Simner, Cuskley and Kirby \(2010\)](#) also demonstrated that associations between linguistic sound and taste exist, with [Cuskley \(2013\)](#) reporting subtle phonological differences in taste related terms across 15 different languages.

Finally, it should be noted that relative iconicity can rely on an arbitrary relationship to depict meaning through the use of certain forms. Examples of this can be seen within phonesthemes, phonemes or phoneme-clusters that are found in small collections of words that share related semantic information ([Bergen, 2004; Otis & Sagi, 2008](#)), such as *gl-*, which can be used to indicate light related properties (e.g. *glow*, *glisten*, *glitter*). Although the relationship between the phonestheme and its semantic information can be considered an instance of non-arbitrariness, given that it carries non-random information which connects certain forms to a meaning, there appears to be no clear link between why a particular phonestheme should represent its associated meaning.

Understanding why sounds such as *gl-* should be used for light related words is unclear, whether it has emerged from a more iconic sound or has always

adopted this apparently arbitrary connection remains a question for future research. Although, recent findings from [Blasi et al \(2016\)](#) have highlighted that instances of relative iconicity, such as the phonestheme *n* to relate to nasal properties, can be observed across thousands of languages, suggesting a more absolute iconic origin to these types of mapping, which given its linguistic universality may be grounded in shared cross-modal associations (see also [Carling & Johansson, 2015](#); [Philps, 2011](#) and [Smith, 2017](#) for promising advances in providing additional explanations on how phonesthemes may have changed diachronically).

One should also note that both forms of iconicity discussed in this section (absolute and relative) may be reliant on using additional cues to enhance their iconicity. The associative strength between an iconic form and its meaning can easily be modulated through the production of the form itself, for instance different prosodic features and gestures can make the iconicity more salient to a listener ([Campisi & Özyürek, 2013](#); [Dingemanse et al, 2016](#); [Perlman, Clark & Johansson Falck, 2014](#)). For example, producing the word *gigantic* to represent something big, can benefit from a deeper pitch and elongation of the sounds to fully exploit the iconic nature of the mapping, however if this word was produced with a high pitch and short duration, the associative strength between the form and meaning can be dramatically reduced. This is something particularly important when one considers the range of acoustic properties human speech can manipulate for expressive purposes ([Nygaard, Herold & Namy, 2009](#); [Shintel, Nusbaum & Okerent, 2006](#)).

Likewise, many studies exploring the effects of iconicity have done so by using 2-alternative-forced-choice designs, with participants asked to map an unfamiliar word form to one of two contrasting meanings (such as in Köhler's 1929 experiment). Although these effects appear robust in laboratory based experiments, their findings are likely more diluted in real world environments, where contrastive referents are not always readily available. Therefore when assessing the strength and validity of iconic mappings, one should also take into account the context in which they are produced (Childs, 1994).

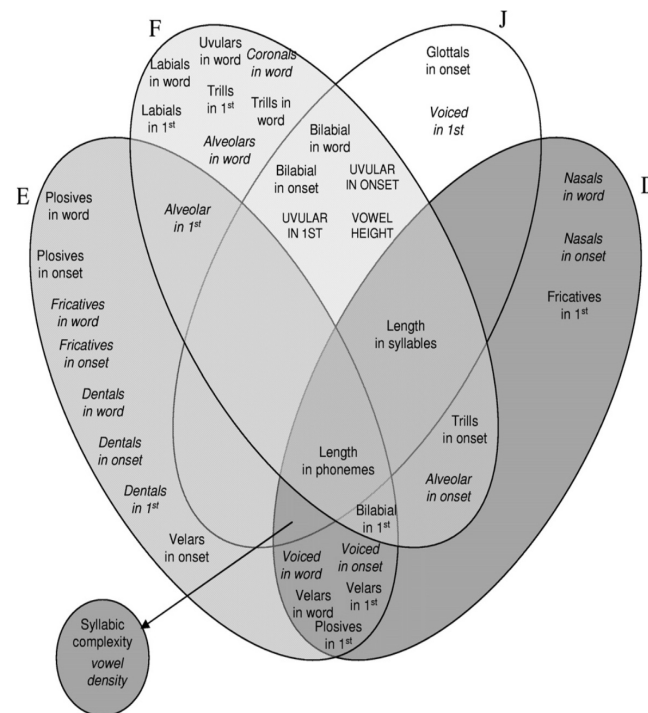
### 1.2.3 Systematicity

Non-arbitrariness in natural language can extend beyond examples of iconic mappings. Systematicity is a form of non-arbitrariness where underlying statistical regularities in a language can be used as cues to distinguish between different groupings of words. Operating on a much larger scale than iconicity, systematicity offers evidence for an inherent structure within the vocabulary, where non-random relationships exist between properties of a word's form and a particular categorization criteria (Tamariz, 2004), see Figure 1.1c. Simple, and uncontroversial, instances of this can be seen within inflectional morphology, for example in English where the addition of *-s* to a root morpheme such as *cow*, changes the word's meaning from a singular form to a pluralised version, i.e. *cows* (Bybee, 1985). This type of non-arbitrariness provides much more general information about the meaning of a word, unlike iconicity where more specific features of meaning are conveyed (Onnis & Christiansen, 2008).

However, systematicity has been reported in many other parts of language outside of a language's morphology. Methodological advances have allowed researchers to explore linguistic corpora from many different languages, which have uncovered a range of subtle cues that can be used to identify important information about a word. For instance, within the vocabulary of several languages, phonological properties, such as phoneme length, are found to be used differentially across certain syntactic categories, e.g. open class words typically have longer phoneme lengths (Kelly, 1992; Farmer, Christiansen & Monaghan, 2006; Monaghan, Chater & Christiansen, 2005, see also Christiansen & Monaghan, 2016 for review).

Interestingly, there is evidence for different languages to adopt different cues. Monaghan, Christiansen and Chater (2007) analysed child-directed speech corpora from four different languages, assessing the relationships between phonological and distributional properties of words with their grammatical categories (see Figure 1.2). They found that although some properties were cross-linguistically reliable predictors of categories, such as the presence/absence of bilabials to make noun/verb distinctions, one could easily assume that there exist languages that do not adopt this relationship. In fact, although many significant relationships were found in Monaghan et al's (2007) analyses, there is no clear evidence that systematicity operates on a perceptuomotor basis, as is the case for iconicity.

This highlights an important feature of systematicity, there does not need to be any clear reason for why a particular property is used to distinguish between



**Figure 1.2.** Venn diagram of phonological cues reported to distinguish between open and closed class words in English (E), Dutch (D), French (F) and Japanese (J). From [Monaghan et al \(2007\)](#).

categories of words (there is nothing about bilabials that should make them suitable candidates to represent nouns and not verbs). The fact that systematicity occurs in ways idiosyncratic to a specific language, could explain why “*le signe est arbitraire*” ([de Saussure, 1916, p.100](#)) has dominated views of language. Instead, systematicity has been proposed as a fundamental structural property of language, with its presence emerging as a result of pressures introduced by the language learner and user ([Christiansen & Chater, 2008](#), [Monaghan et al., 2011](#)). Thus, although the statistical regularities in a language, which provide a non-arbitrary link between form and meaning through systematicity, may differ across

languages and do not incorporate any direct features of meaning, it is a property which pervades the languages of the world (Dautriche, Mahowald, Gibson & Piantadosi, 2016).

### **1.3 Division of labour in the developing vocabulary**

In the previous section, arbitrariness and non-arbitrariness in natural language were introduced and conceptualised, presenting a view of language which acknowledges the presence of both these types of relationships to map form onto meaning. Following on from this view, this section will evaluate the functional properties of arbitrariness and non-arbitrariness in language, with the aim being to address when and why such relationships should influence the structure of the vocabulary.

Languages are acquired, therefore one must learn to map thousands of individual forms onto their respective meanings throughout life. Likewise, languages are used, therefore one must also produce, process and comprehend such mappings, often in real time, in daily life. Thus, a language must i) be learnable and ii) be usable. We will focus on how these two factors act as pressures on the linguistic system, determining the changing influence of arbitrariness and non-arbitrariness in language.

#### **1.3.1 Learnability**

Scholars interested in language have been confronted with a fundamental question: if words offer no clear relationship between form and meaning, and



these words must be learnt, then how can an infant acquire such a system with apparent ease? This question appears more complex when one considers an alternative system that language could adopt, one which is non-arbitrary. In a non-arbitrary system, there would be supplementary information incorporated within the form itself to assist the learner in acquiring language. Indeed, recent empirical investigations have sought to uncover the potential learning benefits that non-arbitrary mappings can provide.

Consider the challenge faced by an infant learning a language. Not only do they have to i) learn what meaning is being referred to by a word form, but also ii) that the word form itself is in fact referential. Non-arbitrariness has been proposed as a tantalizing aid in helping infants overcome these challenges ([Imai & Kita, 2014](#); [Spector & Maurer, 2009](#)). This is because a non-arbitrary mapping can provide a more direct link between form and meaning, which the infant can exploit to establish what is being referred to. Additionally, a non-arbitrary mapping can act as valuable common ground between the infant and the person producing the words, something deemed crucial for establishing a communication system ([Clark, 1996](#); [Scott-Phillips, Kirby & Ritchie, 2009](#)), which could help the infant in understanding that words are not simply just meaningless sounds.

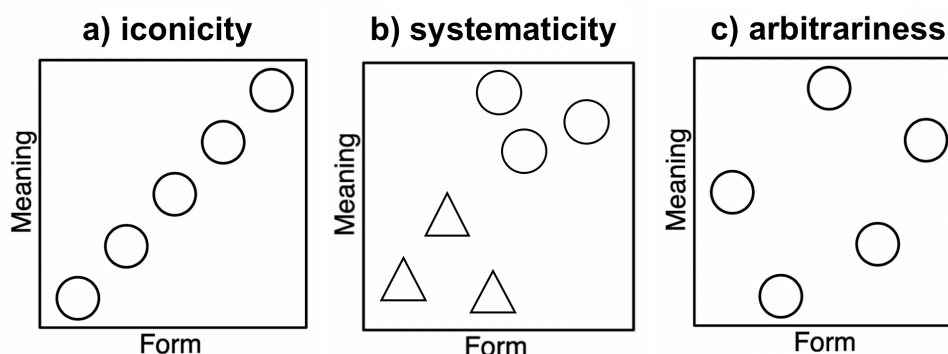
The ability to detect iconic mappings in adults has been well established (see [Lockwood and Dingemanse, 2015](#) for review). Importantly, such sensitivity has also been reported in a range of laboratory based studies with infants (for adjectives: [Peña, Mehler & Nespor, 2011](#); for nouns: [Maurer, Pathman &](#)

Mondloch, 2006; for verbs: Imai, Kita, Nagumo & Okada, 2008). Results from these studies appear to indicate that infants are able to use iconic relationships to identify a referent, with performance reported to be above chance levels, again providing evidence for a non-arbitrary boost for the learner. Such a boost could be highly important for an infant acquiring language, as it may function to help the infant learn that sounds refer to things in the world around them, which in turn could provide a bootstrapping effect for further language development (Imai & Kita, 2014). Although, it should be noted that many studies reporting a learning advantage for iconic mappings often do so when contrasting performance with incongruently iconic mappings, thus the claims may not necessarily extend to an advantage over arbitrary mappings.

Evidence for a non-arbitrary learning advantage has also been reported using computational modelling. Gasser (2004) examined differences in accuracy of learning between an iconic and an arbitrary language system, with an additional manipulation of vocabulary size. The results showed that for an iconic language that had a small number of mappings in its vocabulary, the model outperformed an arbitrary language with the same vocabulary size<sup>2</sup>. Gasser explains this result in terms of language acquisition “We would expect iconicity in circumstances where the number of words is unusually small... The number of words is small early in first language acquisition” (p. 438). This is because when there are few mappings to acquire, the learner can benefit from the neatly ordered

---

<sup>2</sup> See section 1.3.2 for further discussion of the results of this model when the vocabulary size was increased.



**Figure 1.3.** Iconicity, systematicity and arbitrariness in a small vocabulary size (adapted from Gasser, 2004). **a) iconicity**, where a highly correlated relationship between form and meaning exists, enabling easier acquisition of individual words. **b) systematicity**, where similar forms are grouped together, enabling learning of categories (indicated by the shapes, e.g. triangles are nouns and circles are verbs). **c) arbitrariness**, where there is no relationship between form and meaning, making acquisition difficult.

relationship between form and meaning, which iconicity offers, but arbitrariness does not (see Figure 1.3 a and c).

Fundamental support for claims linking processes of acquisition to iconic mappings found in natural language comes from results obtained in a series of experiments by Perry et al (2015). They collected ratings from English and Spanish adults participants, who were asked to judge the iconicity of approximately 600 words. A key finding from the studies revealed that there was a significant relationship between the iconicity ratings and the age at which those words are acquired<sup>3</sup>, with similar findings also being reported in sign language (Vinson et al., 2008). Such results point towards a changing role for iconicity in

<sup>3</sup> See also Perry et al (2016) where the number of words rated was increased substantially to 1,952, but only for English. A significant relationship was again reported for the data.

language development, where iconic mappings appear to populate the language during the earliest stages of acquisition, but then this dominance decreases as the infant's vocabulary develops and the number of words to learn increases.

Elsewhere, systematicity has also been shown to be advantageous for the learner. Given that grammatical categories of words can be distinguished based on the phonological properties of words (Monaghan et al., 2005), researchers have been interested in whether this systematicity can help the learner solve syntactic problems. Indeed, experimental evidence has demonstrated that young children are able to discriminate between words which contrast in their grammatical categories, when the words incorporate the phonological and acoustic cues present in the language they were acquiring (Cassidy & Kelly, 2001; Fitneva, Christiansen & Monaghan, 2009; Shi, Werker & Morgan, 1999).

Similar to iconicity, such systematicity has been reported to be found in the words acquired earliest in life. Monaghan et al (2014) analysed a corpus of monosyllabic English words for systematicity<sup>4</sup>, reporting a significant relationship between systematicity and age of acquisition, with systematicity found to be most prevalent in words acquired earliest, but becomes less pronounced in words acquired later in life. These results suggest that the vocabulary is structured in a way that enables the learner to take advantage of the benefits of systematicity when the vocabulary is relatively small but less so as it develops and expands, again supporting Gasser's (2004) model, see Figure 1.3b.

---

<sup>4</sup> Here, Monaghan et al. measure systematicity by comparing word form similarity to meaning similarity, to examine whether words that sound similar have similar meanings, following Monaghan et al's (2010) methods.

### 1.3.2 Usability

So far, a view of the vocabulary has been presented which considers the challenges facing the language learner and how non-arbitrariness has been suggested to provide specific advantages over arbitrariness. However, if we consider the dynamic nature of language and the fact that it also has to be communicatively functional, not just learnt, then concluding that a non-arbitrary vocabulary offers an optimal system for all of language's demands would be somewhat short-sighted (as suggested by [Dessalles, 2008](#)). Interestingly, attempts at designing a completely non-arbitrary language, and then successfully implementing it, have proven to be inefficacious, with [Wilkins' \(1668\)](#) attempt highlighting some of the shortcomings that a fully non-arbitrary language can create. In his language, [Wilkins](#) attempted to create a system whereby individual letters are used to represent a specific categorical meaning, e.g. animals by the letter *z*, with combinations of these letters being used to provide the general meaning through the word's form.

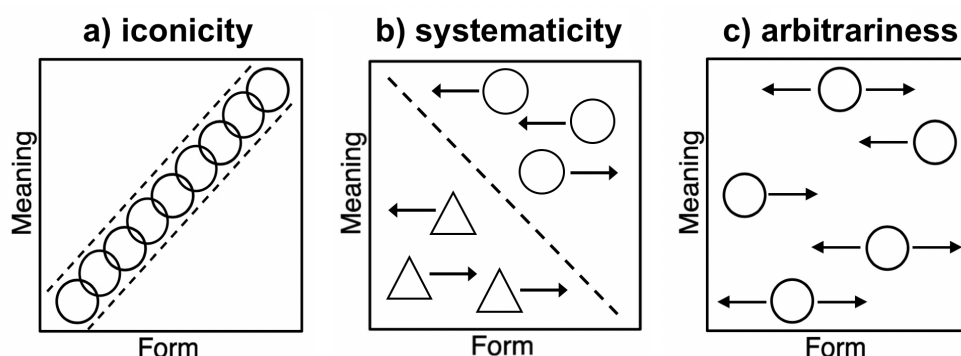
If we consider this language from the viewpoint of a naïve learner, then perhaps the acquisition of a small number of these words, where the meaning categories are relatively distinctive (e.g. to represent an animal (*z*), that is a bird (*e*) and is small (*m*), the word form *zem* could be used), would be plausible. However, when the number of words increases, the number of possible meaning categories increases too, presenting the language user with considerable problems. For instance, a demand for expressivity, where more individuated meanings are conveyed, would result in word lengths being increased, because

more letters would be required to express the intended meaning. When comparing this to arbitrariness, where the flexibility in word forms allows for greater expressivity, a language system such as the one Wilkins proposes would appear much less efficient for both production and processing.<sup>5</sup>

Similarly, a fully non-arbitrary system may also introduce ambiguity in expression, with similar meanings being expressed by similar forms. For instance, if we imagine two distinct word forms required to represent *pigeon* and *dove*, two meanings that share a lot of semantic information. In an arbitrary system, distinguishing between these two meanings is relatively simple as the two words differ greatly in their phonological representations. However, in a non-arbitrary system, this is problematic as the two meanings would have to be represented by forms that share some properties also. In an iconic system, this would mean using two very similar (if not identical) sounds, or if we follow Wilkins' proposed system, this would require additional complexity to avoid ambiguity, leading to confusability between forms. Indeed, as [Eco \(1995\)](#) points out, Wilkins himself encounters such problems by confusing the words for barley and tulip given their similarity (Gade and Gape respectively in Wilkins' language).

---

<sup>5</sup> Although it has been demonstrated that word lengths reflect conceptual complexity, suggesting that a small amount of non-arbitrariness in language can be beneficial to the language user, where longer words actually represent more complex meanings ([Lewis & Frank, 2016](#)).



**Figure 1.4.** Iconicity, systematicity and arbitrariness in a large vocabulary size (adapted from Gasser, 2004). **a) iconicity**, where a highly correlated relationship between form and meaning exists, such a constraint increases ambiguity and overlap of forms. **b) systematicity**, where word forms are relatively flexible (indicated by the arrows), but incorporate a phonological cue to aid processing of categories (indicated by the shapes, e.g. triangles are nouns and circles are verbs). **c) arbitrariness**, where word forms are unconstrained (indicated by the arrows) allowing them to map onto any meaning, providing an efficient and flexible structure to the vocabulary.

Such issues are highlighted in Gasser's (2004) model, where non-arbitrariness is advantageous for a small vocabulary size, however as the size increases this advantage is lost, with arbitrariness offering a more optimal system in comparison. Gasser's results can be explained in terms of a communicative pressure for expressivity, where a large vocabulary is important for clear, accurate and mutually understandable communication (de Boer & Verhoef, 2012; Kirby et al., 2015). If the language system is to adapt in order to meet the demands of this pressure, then an iconic system will encounter similar issues experienced by Wilkins (1668). Specifically, there is much more potential for ambiguity in expression, as similar forms will be used to represent similar meanings, because iconicity is restricted by a correlated relationship between form and meaning, see Figure 1.4a.

Such issues are resolved to an extent by an arbitrary system, which is not restrained by an inherent relationship between form and meaning. The flexibility afforded to an arbitrary system allows for any word form to map onto any meaning, so that words that are similar in meaning are not necessarily related by form. For example the words *pigeon* and *dove* are strongly related by meaning, however their forms are sufficiently distinct from one another to allow discrimination and efficient processing of the intended referent. Likewise, the words *dove* and *glove* are phonologically closely related, yet their meanings are very distinct, see [Figure 1.4c](#).

Yet an arbitrary system alone might not provide the language with optimal conditions for use, with systematicity also playing an important part in the communicative efficiency of a language. There is strong evidence to suggest that there is a processing advantage for words that exhibit greater systematicity between form and meaning, with results from lexical decision tasks ([Monaghan et al., 2010](#)) and word naming latencies ([Farmer et al., 2006](#)). Thus, we are able to capitalise on underlying phonological cues that provide information about a word's meaning/category, so when expanding the vocabulary and responding to pressures of expressivity, we can benefit from the pre-established non-arbitrariness provided by systematicity, whilst also incorporating arbitrary properties simultaneously, see [Figure 1.4b](#).

Perhaps the most robust evidence for this division of labor between arbitrariness and systematicity has been shown by [Monaghan et al \(2011\)](#). Through a series of computational, behavioural and corpus methods, they



reported that when mappings are entirely arbitrary, the language benefits from meaning individuation, where individual word forms can unambiguously map onto individual meanings . Whereas, a systematic language benefits from categorical distinctions, where the word form incorporates a cue that aids the identification of meaning categories. Critically, when the language adopts word forms that incorporate properties of arbitrariness *and* systematicity, then this provides the optimal balance between individuation and categorical learning. Indeed, corpus analyses have demonstrated that this division of labor is incorporated within the structure of words in the vocabulary. [St. Clair, Monaghan and Ramscar's \(2009\)](#) analyses of a child directed speech corpus, revealed that words used suffixes to encode information about grammatical category, whilst [Monaghan et al \(2011\)](#) demonstrated that information relating to the word's individual meaning was located most strongly at the beginning of the word.

## 1.4 Cognitive perspectives on lexical evolution and change

Theories surrounding the origins and evolution of language have been topics for contemplation for many years, with [Darwin \(1871\)](#) postulating “*I cannot doubt that language owes its origin to the imitation and modification, aided by signs and gestures, of various natural sounds, the voices of other animals, and man's own instinctive cries*” ([p. 56](#)). Yet, only since the 1990's has the scientific study of the origins and evolution of language really made empirical progress. One notable theory has been the claim that as language first emerged, it would have

been heavily reliant upon the articulation of meanings through iconic representations (Cuskley, 2013; Ramachandran & Hubbard, 2001). Although such a theory will always remain speculative, considerable evidence has been accumulated, which provides substance to the claims.

Much of the evidence in support of the theory of an iconic origin for human language has stemmed from the advantages it provides to the learner acquiring a language. As discussed in the previous section, iconic mappings appear to be beneficial when one is establishing a shared communication system, as an iconic mapping provides information about meaning and its referential intent. This has led researchers to extend the claims which link iconicity and ease of acquisition, to the establishment of a communicative system used by our ancestors. Although the evidence is indirect, many researchers accept its plausibility<sup>6</sup>, whereby an originally iconic system was used to bootstrap language, but has since evolved into the predominantly arbitrary system of modern language. Again, researchers are steadily building evidence in support of the theory, as demonstrated through experimental communication games (Garrod et al, 2007), as well as rare cases of naturally emerging sign language (Senghas, Kita & Özyürek, 2004). Yet there is the potential for establishing more clear links between acquisition and evolution, to provide insight into the origins of language.

A dominant view of how languages evolve comes from the cultural transmission of language itself. In this process, a language system is acquired

---

<sup>6</sup> Hurford (2011) in particular provides an extensive and eloquently written evaluation of theories on the origins of language.

(normally by an infant), which is then modified slightly in response to cognitive pressures that the next generation of learners will learn from, and over time this combination of acquisition and use leads to adaptations arising (Christiansen & Chater, 2008). Whilst many researchers have made significant progress in identifying how certain properties and structures within a language driven languages to change and evolve, fewer attempts have been made to explain why certain features have remained stable over the course of transmission (Croft, 2000).

One promising innovation in the field which aims to address such an issue has been the application of cladistics, whereby the phylogeny of languages are reconstructed over thousands of years. Using such an approach has enabled researchers to quantitatively trace back the common ancestors of modern languages, building on the work of historical linguistics, on the basis that languages have gradually diverged over the course of many generations (Gray & Atkinson, 2003). Despite this divergence, certain features found in one language appear to persist in other languages, suggesting that such features have been resistant to change over the course of transmission, offering valuable insights into the deep ancestry of languages (Pagel et al., 2013).

Focusing on the evolution of the vocabulary, one non-trivial claim is that certain words undergo change more rapidly than others. By a simple examination of different words for the number ‘2’ across Indo-European languages – *dos*, *deux*, *dwa*, *δύο* – one can observe striking similarities, yet for other words, such as *candle* – *vela*, *bougie*, *świeca*, *κερί* – one now sees striking differences.

Explanations for changes in the vocabulary have normally been offered by social and demographic factors (Labov, 2001), which have undoubtedly provided substantial contributions to our understanding of lexical change. However, more recently explanations for processes of change have been provided by cognitive factors, establishing a link between language change and language evolution (Hruschka et al., 2009).

One key (and well replicated) finding that ties cognitive factors to lexical change, is that the rate at which words change can be predicted by frequency of occurrence (Pagel, Atkinson & Meade, 2007). Pagel and colleagues calculated a measurement for the likelihood of a word being replaced by a new non-cognate form (see Pagel & Meade, 2006 for full methodology). They were then able to accurately predict the rate at which a word underwent lexical replacement based on the frequency of occurrence for that word. The results demonstrated that higher frequency words underwent change less rapidly than low frequency words, with the same relationship found in four different languages. In order to explain the findings, Pagel et al (2007) propose that increased usage prevents a word from processing and production errors, thus making it more resilient to large changes, such as replacement. Such a claim is supported by the vast literature on frequency effects in language (Ellis, 2002), but also more recently by Hay et al (2015), who demonstrated that changes in the pronunciation of words is much more likely when a word is lower in frequency, even over a period of 130 years.

Developing on the idea that psycholinguistic properties can be used to predict the rate of lexical change, [Monaghan \(2014\)](#) proposed that the age at which words are acquired should also be a valid predictor for rate of change, in light of previous evidence showing processing and production advantages for early acquired words ([Catling & Johnston, 2009](#)). Indeed, [Monaghan's](#) analyses revealed that early acquired words were more likely to resist dramatic changes over time, whilst words acquired later were more vulnerable to replacement, even when controlling for various other factors known to predict rates of change, such as frequency and word length. This finding, although only an initial insight, could have significant implications for understanding language evolution more generally. If words are more conserved in the vocabulary when they are acquired early, and therefore undergo less modification, then this places the locus of change on the mature language user, not on the naïve language learner, presenting an alternative account to previous views (for example, [Bickerton, 1990](#)).

Similarly, [Blasi et al \(2016\)](#) demonstrate that within lists comprising basic vocabulary items, certain speech sounds are used to represent certain meanings, and that these associations are found persistently across thousands of unrelated languages. This suggests that basic fundamental items in a human's vocabulary have incorporated sounds which may potentially reflect an underlying cognitive bias akin to iconicity. Furthermore, it would appear that these associations are not only widespread across the world, but they could have actively been retained in the language over the course of transmission across many generations, which

would be consistent with cognitive explanations for rates of linguistic change (Monaghan, 2014; Pagel et al, 2007).

Critically, this provides support for the claim that the vocabulary is structured dynamically, utilising non-arbitrary mappings during acquisition, when learnability is important, but then arbitrary mappings when the pressures of communicative use take precedence. If the acquisition of language benefits greatly from non-arbitrariness in the vocabulary, then these mappings should be preserved as a priority, in order to provide the same benefits for the next generation of learners. In contrast, words acquired later have to be communicatively effective, as a result of a much denser vocabulary, therefore any lexical changes may serve to enhance communication, by potentially maximising discriminability between similar meanings, or even conforming to pre-existing patterns of systematicity in the language.

## **1.5 Thesis Outline**

Within this introductory Chapter, key concepts regarding the structure of vocabulary have been introduced, raising fundamental questions about the nature of mappings in the vocabulary and whether arbitrariness and non-arbitrariness provide solutions to some of the pressures faced by the learner and the user of a language. Additionally, a cognitive explanation for processes of linguistic evolution and change has been presented here, whereby it has been argued that specific processing constraints, introduced by the learner and user, have contributed to the stability and instability of items in the vocabulary. In the

following Chapters, we will seek empirical evidence for these questions and claims, adopting experimental methods which look to address the core aims of the thesis. In Chapter 2, we will test the varying effects of sound-symbolism in a growing vocabulary, Chapter 3 will develop on this, examining more closely the effects of sound-symbolism and arbitrariness in vocabularies of different sizes. Chapter 4 will test the way that psycholinguistic properties of words can influence the type types of lexical change in the vocabulary, whilst Chapter 5 will explore these effects through the cultural evolution of language.

# Chapter 2

## The Changing Role of Sound-Symbolism for the Growing Vocabulary

---

*The work in this Chapter has been submitted to the journal Cognition in March, 2017 and was produced in collaboration with Padraic Monaghan and Peter Walker as co-authors.*

### 2.1 Introduction

The relationship between a word's form and its meaning has long been considered to be arbitrary (De Saussure, 1916; Hockett, 1960). Whilst the vocabulary that an adult acquires largely comprises arbitrary words, recent interest in the presence of non-arbitrary form-meaning mappings has challenged this traditional view that arbitrariness should be considered a design feature of language (Dingemanse et al., 2015). Perhaps the most well documented example of a sound-symbolic relation between form and meaning is the 'bouba-kiki'



effect (Köhler, 1929, 1947; Ramachandran & Hubbard, 2001), where a specific preference is observed for matching particular sounds in non-words with either rounded ('bouba') or spiky ('kiki') shapes (Bremner et al., 2013; Cuskley, Simner & Kirby, 2017; Dingemanse et al., 2016; Kovic, Plunkett & Westerman, 2010; Maurer, Pathman & Mondloch, 2006; Ozturk, Krehm & Vouloumanos, 2013; Walker et al., 2010).

Sound-symbolism may be particularly useful for assisting in learning word-referent mappings at an early stage of language development. Given that a learner is confronted by the difficult task of determining form-meaning mappings (Harnad, 1990; Quine, 1960), sound-symbolism may assist children to learn that words have reference because of an inherited understanding of cross-sensory correspondences (Imai, Kita, Nagumo, & Okada, 2008; Imai & Kita, 2014; Kantartzis, Imai, & Kita, 2011; Maurer et al., 2006; Nygaard, Cook, & Namy, 2009). Thus, learners are provided with information about the meaning of the word by incorporating signification within the actual form used, enabling the learner to realise that the form is potentially referential and, further, what the referent actually is (Ramachandran & Hubbard, 2001; Spector & Maurer, 2009).

The importance of sound-symbolism for early language development is supported by studies of systematicity in form-meaning mappings in the early vocabulary. In an analysis of the vocabulary of English, non-arbitrariness was found to be more prevalent for the words children acquire earlier in language (Monaghan, Shillcock, Christiansen, & Kirby, 2014). For the words children learn between the ages of 2 and 5, there is greater systematicity between form

and meaning of words than expected by chance. Similarly, Perry, Perlman, and Lupyan (2015) found that words that participants rated as iconic, i.e., rated highly as “words that sound like what they mean”, were more likely to be those that children acquire earlier in vocabulary development.

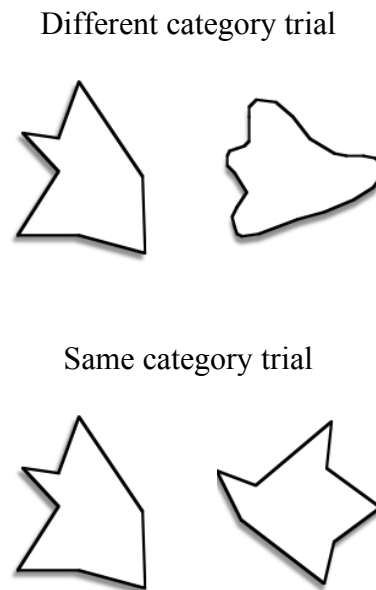
However, the presence of sound-symbolism in the early vocabulary diminishes in the later vocabulary: In Monaghan et al.’s (2014) analysis, from ages 7 onwards, there tends to be greater arbitrariness than expected by chance in form-meaning mappings. Thus, understanding the role of sound-symbolism in language development has to provide an explanation for when sound-symbolism is and when it is not potentially advantageous for learning.

Gasser (2004) predicted that arbitrariness in sound-meaning mappings should be increasingly beneficial for learning as the vocabulary size increases. If word forms contain sound-symbolism then this restricts the possibilities for new words to be interleaved with the representations of previously acquired words, whereas arbitrary relations enable greater flexibility in forming the new word’s mapping. Monaghan, Christiansen, and Fitneva (2011) also predicted from computational modeling that arbitrary relations ought to be advantageous for learning larger vocabularies because they enable a maximizing of the information present in the word-learning environment to be exploited. Thus, sound-symbolism limits the distinctiveness between words with similar meanings which is not problematic when there are just a few words in the vocabulary, but becomes an increasing strain on form-meaning mapping formation as the sound space becomes more populated with a larger vocabulary. However, these

benefits of arbitrariness for learning larger vocabularies over smaller vocabularies has yet to be tested experimentally. Thus, we predict that sound-symbolism is beneficial for learning individual sound to meaning mappings for a small vocabulary, but that this facilitation should reduce with a larger vocabulary.

Though there is increasing arbitrariness at the individual word level for the growing vocabulary (Monaghan et al., 2014; Perry et al., 2015), systematicity at the *category* level is observable across the whole vocabulary. Kelly (1992) showed that there is a systematic correspondence between the sound of words and their grammatical category which applies cross-linguistically (Monaghan, Christiansen, & Chater, 2007). The same idea that phonology can be used advantageously to provide category-level information had driven historic efforts to create entirely systematic, universal languages, whereby meaning could be comprehended simply from the form being expressed (e.g. Wilkins, 1668).

Monaghan, Mattock, and Walker (2012) tested whether learning could be supported by systematicity at the category level. They trained participants to map between 16 non-words and meanings drawn from two shape categories. They varied the extent to which there was a systematic or arbitrary relation between sounds of words and the category distinction. They found that systematicity facilitated learning of the broader category distinctions between words (see also Farmer, Christiansen, & Monaghan, 2006). Thus, though sound-symbolism may be useful for individual word learning for small vocabularies, sound-symbolism ought instead to be beneficial for learning category distinctions for larger vocabularies.



**Figure 2.1.** Examples of a same and different category trial . A congruent mapping would pair a plosive word, e.g., /bIk/ to the angular shape, whilst an incongruent mapping would pair a plosive word with the rounded shape.

In this experiment, we tested the effect of sound-symbolism on learning individual word meanings and category distinctions for different sizes of vocabulary. Adult participants were trained to learn word-referent mappings, where referents were either rounded or angular in shape. Mappings were either congruent with sound-symbolism, where the word was paired with an object to reflect previously established sound-symbolic relations, or incongruent, where the mapping was inconsistent with these relations. Learning trials varied in terms of whether the participant had to discriminate between choices from the two different shape categories (e.g., one angular and one rounded shape were presented), or when the choices were from the same shape category, such that category-level information was not available to support the decision (e.g. both angular), see [Figure 2.1](#).

## 2.2 Experiment 1: Learning from a Congruent and Incongruent Language

### Method

**Participants.** Seventy-two undergraduate students from Lancaster University, with a mean age of 18.7 years ( $SD = 0.8$ , range 17-21) participated. All participants spoke proficient English (55 had English as a first language). Informed consent was collected from each participant and ethical approval was obtained from Lancaster University's ethics committee.

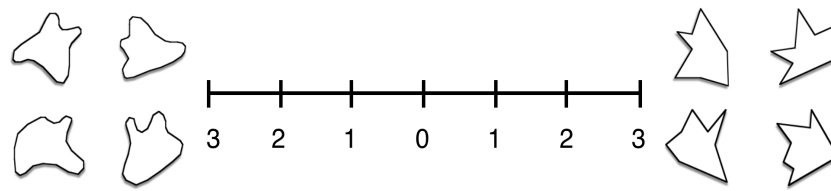
**Materials.** For the visual stimuli, 16 different shapes were constructed which were either rounded or angular in shape (8 shapes for each category). Shapes were similar in terms of perceived size, and complexity in terms of numbers of protuberances (see Monaghan et al., 2012, for details of the controls). For the auditory stimuli, 16 different monosyllabic consonant-vowel-consonant non-words were recorded by a native English speaker in a monotone. For 8 of the non-words, plosives were used for the consonants (/k/, /g/, /t/, /d/, /p/, /b/) in both onset and coda positions. Continuants consisting of nasals, liquids or approximants (/m/, /n/, /ŋ/, /l/, /ɹ/, /w/), comprised the onsets and codas for the remaining 8 non-words. Each non-word contained a vowel chosen from one of the following four sounds (/æ/, /ɛ/, /ɪ/, /ʊ/). Each vowel was used an equal number of times within the sets of rounded and angular non-words. The full list of non-words used can be found in [Table 2.1](#).

**Table 2.1.** List of phonetically transcribed words used during Experiment 1.

Continuant Words	Plosive Words
mɒŋ/	/kɪb/
/nim/	/gæt/
/læn/	/tɛg/
/ɪɛŋ/	/dɒp/
/wɒl/	/pɛd/
/wɛm/	/bɪk/
/ɪn/	/tɒb/
/næl/	/kæg/

To ensure that the sounds used were reliably sound-symbolic, twenty-two additional participants completed a short questionnaire rating the strength with which they felt each sound corresponded to rounded or spiky shapes, which were illustrated on either side of a 7 point scale. The scale consisted of ‘0’ for no correspondence, ‘1’ for weak, ‘2’ for medium, and ‘3’ for strong correspondence, and an example item is shown in [Figure 2.2](#). Ratings indicating an angular shape preference were coded as negative values. Plosive non-words were judged to correspond more closely to angular than rounded shapes (mean rating = -.58, SD = 1.49), whereas continuant non-words more closely corresponded to rounded shapes (mean rating = .18, SD = 1.37), and scores were significantly different,  $t(672.55) = -6.867, p < .001$ .

For the vocabulary learning task, sounds were mapped to the shapes in two different ways for each participant. Half the mappings were congruent with



**Figure 2.2.** Example of Likert scale item for correspondence between word and rounded or angular shapes. Rounded shapes were presented on the left side of the scale for half the trials and on the right for the other half.

previous sound-symbolic studies of phoneme to shape mappings (Fort, Martin & Peperkamp, 2015; Nielsen & Rendall, 2012), where rounded shapes were mapped to the continuant non-words, whilst angular shapes were mapped to the plosive non-words. The other half of the mappings were incongruent, which paired rounded shapes with plosives and angular shapes with continuants. Participants were exposed to an equal number of congruent and incongruent trials during the experiment.

The small vocabulary condition presented 4 rounded and 4 angular images and 4 plosive and 4 continuant non-words, selected randomly from the set of 16 images and 16 non-words for each participant. The medium size vocabulary condition selected 12 images and 12 non-words from the set of 16. The large vocabulary size utilised all 16 images and non-words, and was thus similar in design to Monaghan et al. (2012).

**Procedure.** A cross-situational learning paradigm was used in the experiment (see Smith & Yu, 2008). Participants heard a sound and viewed two shapes side by side on a computer screen, and were required to decide which

shape they thought the sound referred to, pressing “1” or “2” on a computer keyboard to select the left or right shape, respectively. One image was the target, which always co-occurred with the spoken word, and one was the foil, which was one of the other images in the set to be learned. Positions of targets and foils was counterbalanced within blocks of trials, and no feedback was given.

The foil was either a shape from a different or same shape category as the target, in order to test whether a broad categorical distinction, or individual word meanings were learned, see [Figure 2.1](#). Learning is therefore tested by ability to discriminate between two alternatives, which is a standard method for testing word learning (e.g., Horst, Samuelson, Kucker, & McMurray, 2011). There were 4 blocks of training, within which each mapping was presented 4 times. As the number of mappings varied in each vocabulary condition, the number of trials per block also varied: 32 trials per block for the small, 48 trials for the medium, and 64 trials for the large vocabulary condition.

## Results

In the analysis conducted on the data, we modelled the probability (log odds) of response accuracy, accounting for the variation across participants and stimuli. Observations were clustered for each participant and stimulus, therefore we performed a series of Generalized Linear Mixed-effects Models (Baayen, 2008; Jaeger, 2008), specifying first the random effects. Then, we considered the effect of experimental condition (vocabulary size), the effect of block over the course of the experiment, the effect of learning trial type (same or different category presentation) and also the effect of congruency. We then considered the

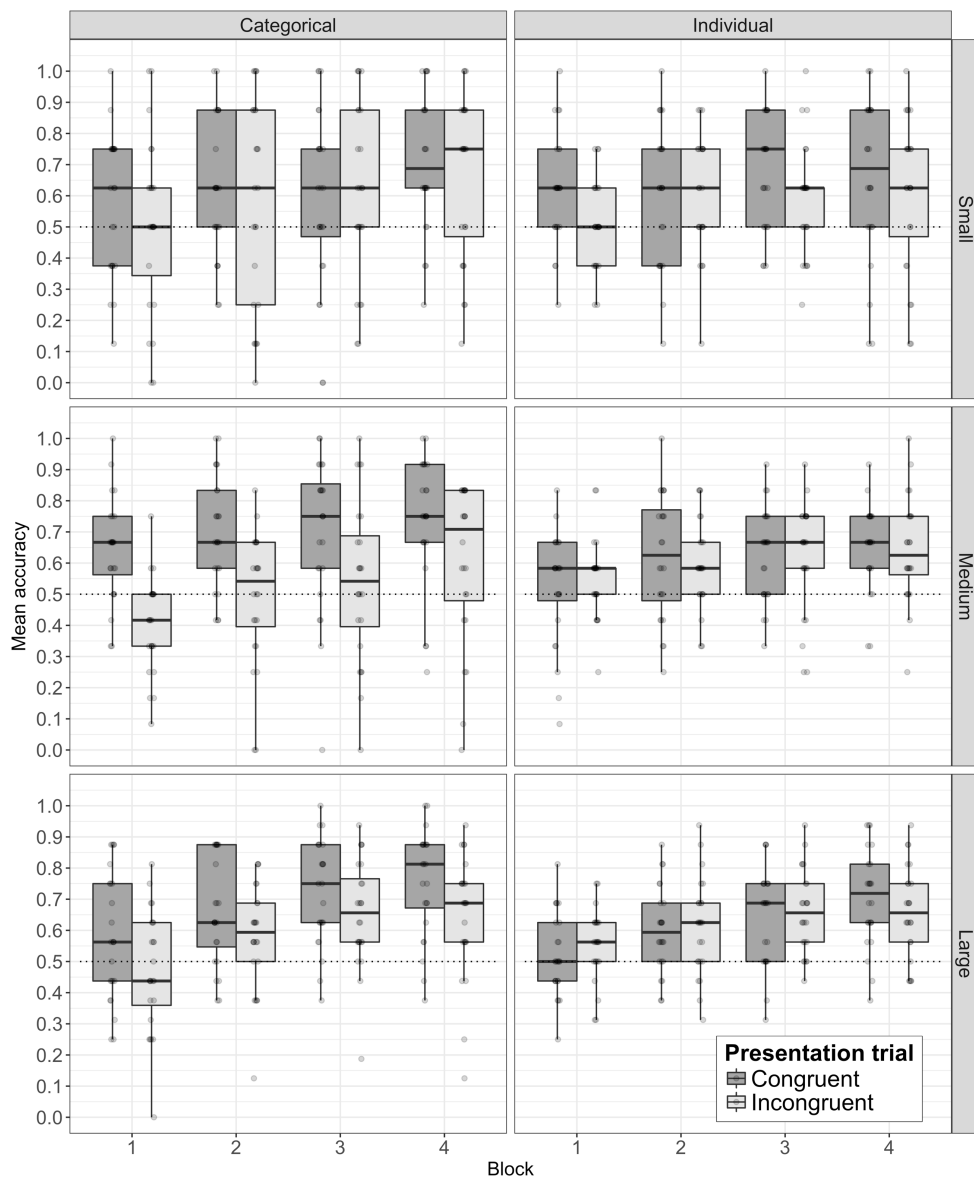


interaction between vocabulary size, same or different shape condition and congruency. After adding each fixed effects to the model, we ran likelihood ratio test comparisons, comparing the new model to the previous one. This showed whether the inclusion of the new term significantly improved the fit of the model.

Adding the effect of vocabulary size to a model with just random effects did not significantly improve the fit of the model,  $\chi^2(2) = .97, p = .62$ . The inclusion of the effect of block significantly improved the fit of the model,  $\chi^2(3) = 153.1, p < .001$ , this effect was found to be significant and positive, indicating that performance over the course of the experiment improved: estimated intercept log odds for the model = .20,  $SE = .02, z = 12.33, p < .001$ , see [Figure 2.3](#).

Additionally, including the interaction term of vocabulary size x congruency x categorical/individual learning also significantly improved model fit,  $\chi^2(8) = 31.5, p < .001$ . This indicated that the effect of sound symbolism for the categorical and individual learning tasks varied as a function of vocabulary size. The interaction was significant in a positive linear fit (estimate = .39,  $SE = .13, z = 2.98, p = .003$ ). Full details of the model selection can be found in [Table A1.1](#) and the final model summary in [Table A1.2](#).

To understand this three-way interaction, we tested models investigating performance for categorical and individual word learning trials separately, allowing us to explore the two-way interactions between vocabulary size and congruency. For categorical trials, the inclusion of the interaction term as both a linear and quadratic effect significantly improved model fit,  $\chi^2(4) = 24.2, p < .001$ . In follow-up one-way analyses, congruency improved model fit for the



**Figure 2.3.** Proportion of correct responses by block, for same and different category presentations, by vocabulary size condition) for Experiment 1. Dots represent individual subject data. Dotted line shows 50% chance level.

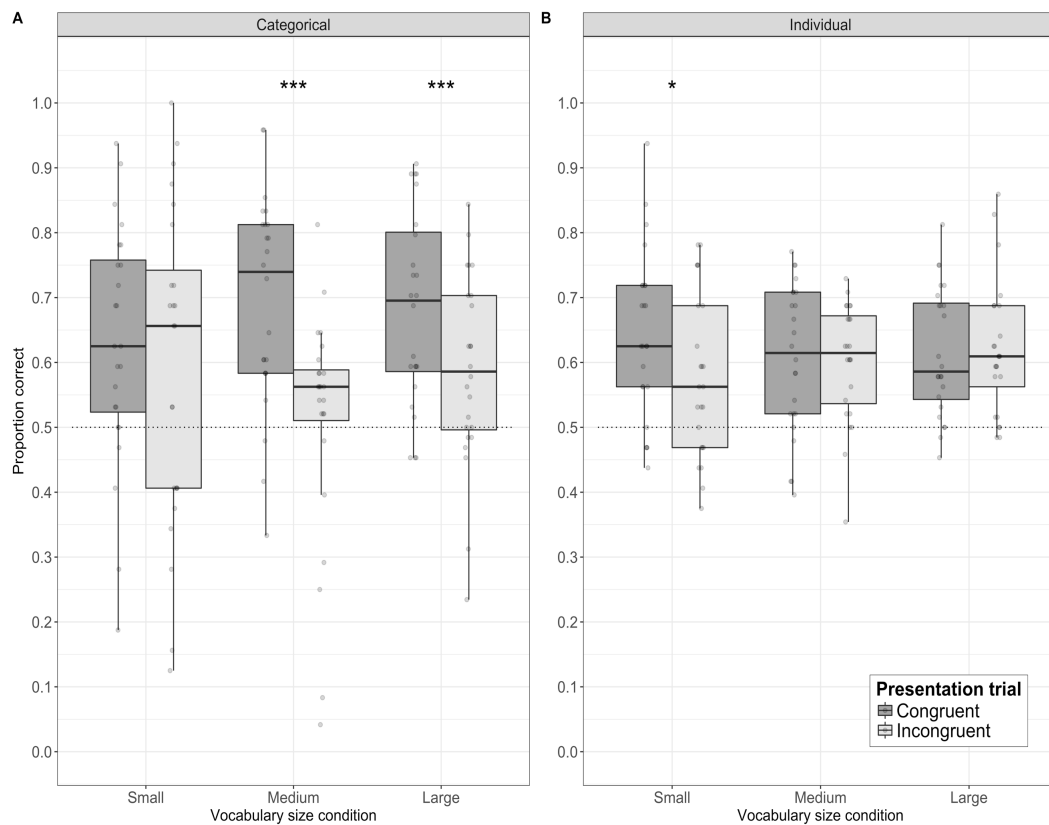
medium and large vocabulary sizes,  $\chi^2(1) = 86.399$ , and  $\chi^2(1) = 30.437$ , both  $p < .001$ , however for the small vocabulary size, congruency did not significantly improve model fit,  $\chi^2(1) = 2.3061$ ,  $p = .13$ , see [Figure 2.4a](#). Thus, sound symbolism boosted categorization only for the medium and large vocabularies.

As there are more items within the category for the medium and large vocabularies than the small vocabulary, this will have had the effect of strengthening category-level sound symbolism effects in these larger vocabularies.

For individual word learning trials, the linear and quadratic interaction terms did not improve model fit,  $\chi^2(5) = 7.5, p = .19$ , although the linear interaction effect was significant in the model,  $p = .017$ . In follow-up one way analyses, whereas congruency improved model fit for the small vocabulary size,  $\chi^2(1) = 6.5879, p = .01$ , for the medium and large vocabulary sizes, congruency did not significantly improve model fit,  $\chi^2(1) = .012, p = .91$  and  $\chi^2(1) = .0561, p = .81$ , respectively, see [Figure 2.4b](#). Thus, sound symbolism promoted learning individual word-shape mappings but only for the small vocabulary.

## 2.3 Discussion

This study demonstrated one of the reasons for why sound-symbolism is evident in early vocabulary development but why arbitrariness is dominant for later vocabulary development (Monaghan et al., 2014; Massaro & Perlman, 2017; Perry et al., 2015). We showed that when the vocabulary is small, as in the first stages of vocabulary acquisition, sound-symbolism is advantageous for learning the meanings of individual words. Thus, sound-symbolism can effectively be incorporated into the vocabulary structure to support acquisition of word-referent mappings (Imai et al., 2008; Kantartzis et al., 2009; Nygaard et al., 2009).



**Figure 2.4.** Proportion of correct responses in (A) different category presentation trials (categorical learning) and (B) same category presentation trials (individual word learning) ) for Experiment 1. Dots represent individual subject data.

\* $p < .05$  and \*\*\* $p < .001$ .

However, for the larger vocabulary sizes, the advantage at the individual word level for sound-symbolism was not observed, instead sound-symbolism was advantageous but only for learning category distinctions. This provides a potential explanation for why vocabulary acquired later in life does not contain sound-symbolism for individual words (Monaghan et al., 2014) but does demonstrate systematicity between sounds and categories of words (Farmer et al., 2006; Kelly, 1992; Monaghan et al., 2007).

These findings highlight the potential benefits to learning that sound-symbolism can provide at different stages of vocabulary development. When a

language learner is initially acquiring a vocabulary, sound-symbolism may provide an effective, even essential, scaffold that aids the acquisition of the first words in the vocabulary (Kantartzis et al., 2011). This may then provide a bootstrapping effect, allowing for a more densely populated vocabulary to be subsequently acquired (Imai & Kita, 2014). For a larger vocabulary, an arbitrary system becomes more suited for the demands of communication, with non-arbitrariness applying only at the level of distinguishing categories rather than individual meanings. Thus, the general processing constraints introduced by a growing vocabulary are reflected in children's vocabulary acquisition. Language appears to be structured to promote sound-symbolic mappings early on in vocabulary learning, but as the vocabulary expands, arbitrary mappings become dominant as the communicative system demands greater expressivity and signal efficiency.

Our demonstration of the changing effects of sound-symbolism as the vocabulary size increases, provides the first behavioural demonstration of predictions revealed by theoretical and computational modelling, highlighting the advantages of arbitrariness for larger vocabularies and sound-symbolism for when the vocabulary is smaller. Our work thus provides an answer not only to the question as to why sound-symbolism is prevalent in early vocabulary, but also why arbitrariness is dominant as the vocabulary increases. We see these questions as related and have provided a single framework, grounded in computational theories of cross-modal mappings (e.g., Gasser, 2004), that identifies the vital role of both systematic and arbitrary mappings in the vocabulary of a language.

We have shown that observations of sound-symbolism being more prominent in early- than late-acquired vocabulary in natural language studies are supported by the learning advantages observed with different vocabulary sizes. This is also consistent with views of the evolution of language, whereby a sound-symbolic system may have been key during a proto-language stage (e.g., Ramachandran & Hubbard, 2001), but as language evolved under communicative pressures of increasing expressivity, arbitrariness comes to dominate the communicative system.

# Chapter 3

## Form-meaning Mappings and the Structure of the Vocabulary

---

### 3.1 Introduction

The aim of this chapter is to present a more comprehensive account of the way that sound-symbolism and arbitrariness work in a dynamic way, by distributing the two types of mappings within a vocabulary, in order to optimize learning. It builds upon the previous chapter, where a language comprising both sound-symbolically congruent *and* incongruent form-meaning mappings were used to explore such differences in learning advantages. Continuing on from this line of investigation, we will present additional evidence for the hypothesis that different vocabulary sizes are better suited to accommodate different types of form-meaning mappings. The findings reported in [Chapter 2](#) provide initial support for the claim that sound-symbolism can act as an aid to learners when acquiring

form-meaning mappings, consistent with previous findings (e.g. [Imai & Kita, 2014](#)). Yet, the results also provided insights into the different types of learning that sound-symbolism can and cannot aid, with emphasis placed on when these learning advantages are influenced by vocabulary size.

For instance, incorporation of sound-symbolic features within a word form has been shown to be advantageous when learning to distinguish between different categories, but when individuated meanings have to be learnt, this benefit is lost ([Lupyan & Casasanto, 2015](#); [Monaghan et al., 2012](#)), with arbitrary mappings being advantageous to the learner when learning the mappings for individual form-meaning mappings ([Gasser, 2004](#); [Monaghan et al., 2011](#)). However, [Chapter 2](#)'s results provide evidence for the claim that there is a division of labour between sound-symbolism and arbitrariness, with small vocabularies being best suited to sound-symbolism, but as the vocabulary grows in size, then the benefits are limited to learning of broad distinctions, consistent with the results from a computational simulation by [Gasser \(2004\)](#), whilst also providing insight into why fully acquired language is not dominated simply by sound-symbolic mappings, but instead effectively incorporates arbitrariness to suit the needs of the users.

Languages develop over time, with the learner gradually acquiring more and more words to build up their vocabulary. This dynamic aspect of language could allow for arbitrariness and sound-symbolism to aid learning at different stages of development, with the organization of the vocabulary potentially reflecting such learning biases. Indeed, recent research has aimed to demonstrate



this empirically, by assessing the extent to which the distribution of sound-symbolic and arbitrary mappings can be predicted on the basis of knowing the point at which the mapping was acquired. Iconic word forms have long been associated with the words produced during infancy, with onomatopoeic word forms being reported consistently among those produced earliest ([Elsen, 1991](#); [Laing, 2014](#); [Tardif et al., 2008](#)). Examples of sound-symbolism outside of onomatopoeia have also been rigorously investigated, with robust reports linking the presence of sound-symbolism (found within communicative representations) with stages of language development. In the domain of sign language for instance, signs that have been learnt at earlier stages of acquisition have been shown to be judged as having more iconic properties than those learnt later ([Vinson et al., 2008](#); [Thompson et al., 2012](#)). This would suggest that earlier acquired signs are purposefully sound-symbolic, with the learner possibly exploiting such non-arbitrariness to acquire the signs with greater ease at this early stage of acquisition.

Within spoken languages however, there has been an increasing interest to find similar patterns of sound-symbolism within the structure of the developing vocabulary. For instance, sound-symbolic mappings have been shown to have processing and learning benefits over non-sound-symbolic mappings for toddlers ([Imai et al., 2008](#); [Kantartzis et al., 2011](#); [Maurer et al., 2006](#); [Ozturk et al., 2013](#)), although these studies only tested the benefits of learning to discriminate across broad categorical distinctions, such as distinctions between antonym pairs (e.g., hot/cold). More recently, the application of novel statistical

methodologies have shed a more comprehensive light on the extent to which non-arbitrary relationships can be found in natural language. In a study by [Tamariz \(2008\)](#), it was demonstrated for a subset of the Spanish vocabulary, there exists a systematic relationship between phonologically similar words and their contextual co-occurrence in speech, highlighting the presence of an underlying statistical regularity in speech, where words that share sounds will be more likely to share meaning, a finding that has since been shown robustly in 100 languages ([Dautriche, Mahowald, Gibson & Piantadosi, 2016](#)). [Tamariz](#) also acknowledged that this systematicity within the language may cause problems for comprehension and production, as too much similarity across similar meanings may lead to problems with discriminability (an issue we will return to). The analyses revealed that additional features, such as vowel stress, act as a pressure to reduce the potential impact of discriminability, thus rendering the vocabulary in a way that responds to pressures for systematicity, but also for communicative clarity. Yet, there was no link made between whether this systematicity was more prevalent in words acquired earlier on in life.

These two pressures were the focus of a computational simulation by [Gasser \(2004\)](#), who aimed to demonstrate that vocabulary size acts as a critical component for when sound-symbolism and arbitrariness in the vocabulary is most effective. By manipulating the number of form-meaning mappings within a vocabulary that contained either sound-symbolic mappings (where the relationship between form and meaning was closely related) or a vocabulary that had arbitrary mappings (where the relationship was random). The sound-

symbolic language showed optimal learning when the vocabulary size was small, as the learner can exploit the benefits of sound-symbolism without issues of meaning discrimination arising, given that the meaning space is sparsely populated. As the vocabulary size increases however, an arbitrary language provides the most optimal system for the learner, as new lexical items could be learnt efficiently without experiencing ambiguity, a primary issue with a sound-symbolic language with a densely populated meaning space (see [Wilkins' \(1668\)](#) attempt at creating a fully sound-symbolic language). If one considers that when the vocabulary is smallest, the learner is in the earliest stages of language acquisition, with a relatively sparse set of form-meaning mappings, this therefore would indicate that sound-symbolism exerts most benefit during these earliest stages, unlike an arbitrary vocabulary that is most beneficial to the learner once the first lexical items have been acquired and the basis of the vocabulary has been established.

The prevalence of sound-symbolism in the vocabulary has only recently been examined in sufficient depth to demonstrate strong evidence that contradicts long held beliefs about the dominance of arbitrariness within language ([de Saussure, 1916](#); [Hockett, 1960](#)). In a large corpus analysis of English monosyllabic words, [Monaghan et al \(2014\)](#) demonstrated higher rates of systematicity than would have been expected by chance in the data, where systematicity was measured in a similar way to [Tamariz \(2008\)](#), where a distance measurement between word forms and meaning was calculated. Importantly though, the presence of systematic mappings was most obvious in words that

have the earliest age of acquisition ratings, with the effect decreasing as age increases. In another study by [Perry et al \(2015\)](#), age of acquisition was found to significantly predict rates of iconicity (defined as a direct relationship between form and meaning, whereby a word sounds like what it means, which is therefore distinct from systematicity<sup>7</sup>) in English and Spanish, even when controlling for other linguistic properties. These findings reveal not only the presence of sound-symbolism and its previously underestimated prevalence in the vocabulary, but crucially that it is found most strongly within the words we acquire earliest in life.

The extent to which sound-symbolism benefits the learner as their language becomes more developed and the vocabulary becomes more heavily populated with form-meaning mappings is also of theoretical importance. In a series of experiments, [Farmer et al \(2006\)](#) demonstrated that phonological typicality, where phonological properties of a word are shared across other words within the same lexical category, actively improves word and sentence processing. Such results could be interpreted as evidence for systematicity working within a large vocabulary to aid the learning of word categories, where the sound of the word provides some statistically reliable information about the lexical category of the word, i.e. if it sounds like a noun, it probably is a noun.

This view of systematicity benefiting categorical learning within a vocabulary is also supported by [Monaghan et al's \(2012\)](#) study, where there was

---

<sup>7</sup> In their study, [Perry et al. \(2015\)](#) measured iconicity subjectively, where participants were provided with an example of model iconicity, such as '*slurp*' and '*teeny*'. Thus, no distinction was made between relative or absolute iconicity in this case.

a benefit for categorical learning of angular/rounded shapes when the forms were sound-symbolically congruent, such that participants would perform better when there was reliable phonological information that could be used to distinguish whether the referent was angular or rounded. Interestingly however, [Monaghan et al's](#) study did not show any benefits of sound-symbolism when learning individual items, i.e., when participants had to distinguish between two referents that were drawn from the same shape category. This suggests that sound-symbolism works most effectively for learning of broad categories, but not the learning of individuated items within the vocabulary. The results presented in [Chapter 2](#) further our understanding of this point, by demonstrating that the benefits related to sound-symbolism for categorical distinctions are observable when there is a large vocabulary, however when the vocabulary size is small, then sound-symbolism aids the learning of individual meanings, as the meaning space is more sparsely populated and ambiguity within the language is minimized, consistent with the model presented by [Gasser \(2004\)](#).

There is however a strong theoretical framework that stresses the importance of arbitrariness within a language system, that works alongside the presence of sound-symbolism. Whilst we have thus far discussed the benefits of incorporating sound-symbolic properties within the phonological space of a vocabulary, the relative benefits of arbitrariness in the same vocabulary are also crucial to the efficiency and expressivity of the communicative system. [Monaghan et al \(2011\)](#) put forward the argument that language has adapted to incorporate both aspects of arbitrariness and systematicity, allowing for a

division of labour within the structure of language that allows for the benefits of systematicity, but also provides a response to pressures for an expressive and efficient system. Evidence is presented by [Monaghan et al](#) that demonstrates arbitrary form-meaning mappings offer the most effective solution to the problem of learning individuated items in the vocabulary, because the arbitrary system allows words to be distinct enough from each other that the learner is not confronted by issues of ambiguity, a problem with a fully sound-symbolic system (see also [Gasser, 2004](#)).

Whilst [Chapter 2](#) presented evidence for division of labor for sound-symbolism and arbitrariness in language learning, there is still the need for additional empirical evidence to generate a more comprehensive explanation of the ways sound-symbolism and arbitrariness are used during vocabulary development. One important consideration about the design used in [Chapter 2](#) was that the language adopted either sound-symbolic mappings that were congruent or mappings that were incongruent. For these incongruent mappings the relationship between form and meaning may not have been entirely arbitrary, as even incongruent mappings will incorporate some systematic information, therefore it could be argued that such a relationship is not in fact random.

For example, the sound-symbolic form-meaning mappings in [Experiment 1](#) were designed so that plosives were incorporated into all the forms being mapped onto angular shapes, thus providing a systematic relationship between form and meaning that is based on a reliable sound-symbolic association. However, when plosives are incorporated into the forms that are mapped onto

rounded shapes (as was the case in the incongruent form-meaning mappings of [Experiment 1](#)), there is still a systematic relationship between form and meaning, as plosives are still being used systematically, but this time to map form and meaning in an incongruent manner. As this systematic relationship is not random, one could argue that even an incongruent set of form-meaning mappings is not entirely arbitrary.

We present here two experiments that explore the way that vocabulary size can influence the learning of categorical distinctions and individuated words within an artificial language that adopts a fully sound-symbolic system, and a system where there is no relationship between form and meaning, representing a fully arbitrary system. This builds upon the work presented in [Chapter 2](#), however we examine here the two systems independently. By modifying the design of [Experiment 1](#), we aim to provide a more valid examination of the way arbitrariness and sound-symbolism might exert different pressures on the learner. We hypothesize that a fully sound-symbolic system will aid categorical distinctions in a larger vocabulary, whilst it will aid learning of individuated meanings in a small vocabulary. For the arbitrary system, it is predicted that learning of individual meanings will increase as the vocabulary size increases.

## **3.2 Experiment 2: Learning from a fully sound-symbolic language**

In [Chapter 2](#) we presented a way to examine how sound-symbolism may facilitate learning of individual items in a small vocabulary, whilst aiding

categorical distinctions in a large vocabulary. In the present experiment, we aim to explore whether these benefits remain when the language only uses sound-symbolically congruent mappings.

## Method

**Participants.** Seventy-two participants took part in the experiment (48 female), with 24 assigned to each vocabulary size condition. This was selected to be the same sample size as in the experiment from [Chapter 2](#). Participants were undergraduate students from Lancaster University, with a mean age of 19.4 years ( $SD = 1.57$ , range 18-26). It was not required that participants spoke English as their first language (English first language speakers:  $n = 50$ ), but all participants spoke English competently. Informed consent was collected from each participant and ethical approval was obtained from Lancaster University's ethics committee.

**Materials.** The same visual and auditory stimuli were used in the present experiment as those used in [Chapter 2](#) and [Monaghan et al \(2012\)](#). This comprised an inventory of 16 visually presented shapes (8 angular and 8 rounded) and 16 auditorily presented non-words (all monosyllabic and had a CVC structure, with 8 non-words generated using plosive consonants and 8 using continuant consonants, see [p.36](#) for more details and [Table 2.1](#) for the full set of non-words used). Each of the non-words was reliably mapped to one of the shapes. Using these mappings, we generated 3 different artificial languages, each of which differed in the number of mappings it used. Again, this was either a



small vocabulary (8 mappings), medium (12 mappings) or large (16 mappings), all with an equal number of angular and rounded shapes, drawn from the inventory of 16 shapes and non-words. Critically, in the present experiment all form-meaning mappings were presented to reflect only a congruent sound-symbolic relationship (unlike [Chapter 2](#), where mappings were presented as both congruent and incongruent). This meant that all sounds with continuant consonants were mapped exclusively to rounded shapes, likewise all sounds with plosive consonants were mapped exclusively to angular shapes. Which shape the sound was mapped to during the experiment was randomly selected from the set of rounded/angular shapes.

**Procedure.** The procedure was similar to that used in [Chapter 2](#) (see [p.39](#)), in that participants were exposed to the form-meaning mappings within a cross-situational learning paradigm. Participants were presented with two images on a computer screen, as well as hearing a non-word over a pair of headphones. After hearing the non-word they had to choose which image the word was referring to, with one of the images being the target and the other a foil. Each mapping was presented 4 times throughout the experiment, over the course of 4 blocks. Trials were designed to assess two distinct learning scenarios: learning of broad categories (by presenting two images that differed in their shape, such as one rounded and one angular), or learning of individuated meanings (by presenting two images from within the same shape category, such as two rounded shapes), see [Figure 2.1, p.35](#) for an example of each trial type.

In contrast to [Chapter 2](#), where congruency was manipulated as an independent variable, all mappings in the present experiment were treated as sound-symbolically congruent. Therefore, the only manipulation of interest is performance during the two different learning trial scenarios (categorical and individual learning). Performance was again measured throughout the course of the experiment, with participants providing an accurate response if they chose the target image during a trial.

## Results and Discussion

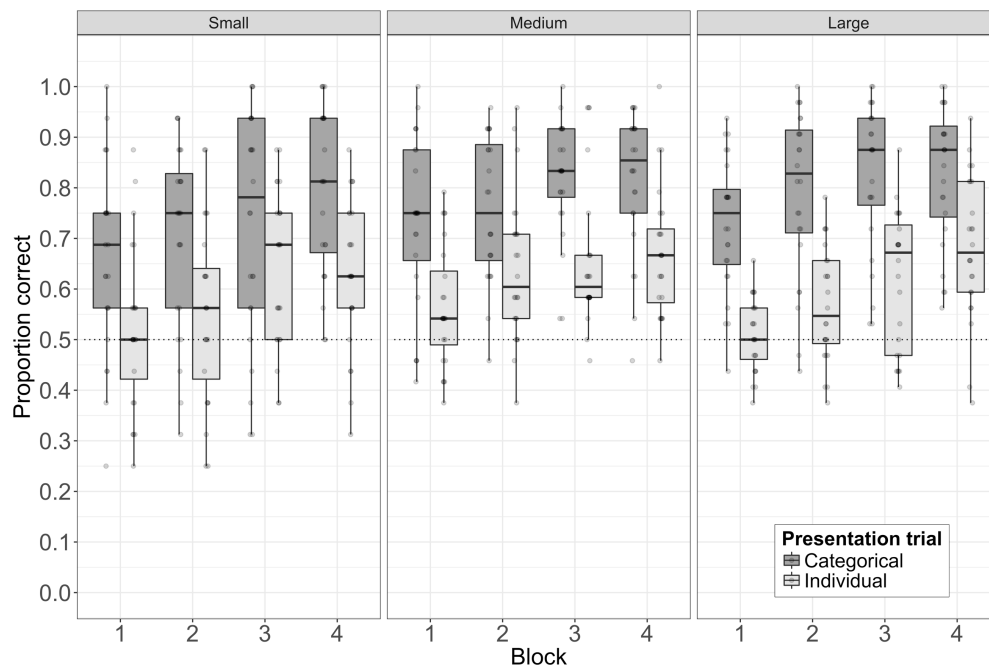
The analysis carried out on the data involved running a sequence of generalized linear mixed-effects models ([Baayen, 2008](#); [Jaeger, 2008](#)), which were fitted to the data to predict response accuracy. In order to account for variation across participants and items, we first specified random effects of subject and target stimulus. We then proceeded to build up the complexity of the model through the addition of fixed effects, running Likelihood ratio tests after a new fixed effect was added. This allowed us to assess whether the additional complexity introduced to the model was justified, with significant effects being included in subsequent models (see [Barr, Levy, Scheepers & Tily, 2013](#)).

The inclusion of the block effect significantly improved model fit ( $\chi^2(1) = 136.36, p < .001$ ), indicating that accuracy increased significantly over the course of the experiment. Further analyses using one-sample t-tests showed that for all experimental conditions, performance was significantly above the 50% chance level at the last block of the experiment (see [Figure 3.1](#) for results). The inclusion

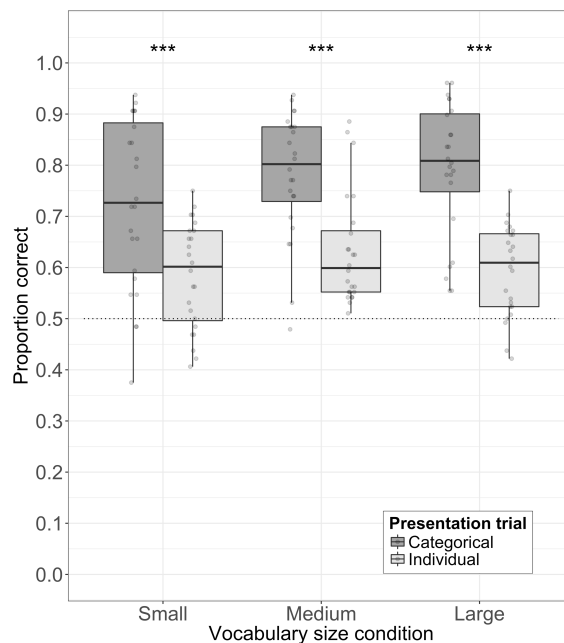
of vocabulary size did not significantly improve model fit ( $\chi^2(2) = 2.570, p = .278$ ), indicating that there were no significant differences in accuracy for the different vocabulary size conditions. There was a significant improvement to model fit when presentation type (categorical or individual) was added, ( $\chi^2(1) = 500.930, p < .001$ ), indicating that accuracy was significantly higher for trials that tested categorical learning, compared to trials testing individual word learning.

Next, the inclusion of the interaction term vocabulary size x presentation type significantly improved model fit ( $\chi^2(4) = 17.529, p = .002$ ). In line with our hypothesis, this indicates that there was a significant difference in accuracy for different vocabulary sizes as the presentation type (categorical or individual) varied. Additional analyses on the separate vocabulary size conditions, revealed that this difference between categorical and individual word learning was present in all three vocabulary sizes (small:  $\chi^2(1) = 73.394, p < .001$ ; medium:  $\chi^2(1) = 137.73, p < .001$ ; and large:  $\chi^2(1) = 310.94, p < .001$ ), with no interaction with the effect of block (all  $p$ 's  $> .05$ ), see [Figure 3.2](#) for results.

However, follow up analyses investigating accuracy on the categorical and individual learning trials separately, revealed no significant improvement to model fit when vocabulary size was added as a fixed effect (categorical:  $\chi^2(2) = 1.998, p = .368$ ; individual:  $\chi^2(2) = 3.103, p = .212$ ). For both categorical and individual learning trials, the results presented in this experiment demonstrate a clear advantage for learning to distinguish between distinct categories, when the language comprised only sound-symbolic mappings. This is consistent with previous reports that



**Figure 3.1.** Proportion of correct responses by block, for same and different category presentations, by vocabulary size condition) for Experiment 2. Dots represent individual subject data. Dotted line shows 50% chance level.



**Figure 3.2.** Proportion of correct responses in different category presentation trials (categorical learning) and same category presentation trials (individual word learning) for Experiment 2. Dots represent individual subject data. \*\*\* $p < .001$ .

suggest sound-symbolism, or more specifically systematicity, facilitates the learning of broad categorical boundaries (Farmer et al., 2006; Lupyan & Cassasanto, 2015; Monaghan et al., 2011, 2012).

Interestingly, this facilitative effect was observed in all the vocabulary size conditions, suggesting that when a language adopts a sound-symbolic vocabulary, there is always an advantage when learning to distinguish between different categories of meaning in the vocabulary. Although in Chapter 2, the results showed no significant difference between the sound-symbolically congruent and incongruent mappings for learning of categorical distinctions, when the vocabulary size was small, the results presented here may indicate that this facilitative effect may only be observable when the mappings incorporate a largely sound-symbolic vocabulary. Given that the language presented in Chapter 2 utilised both congruent and incongruent mappings, the incongruent mappings may be considered to hold some systematicity within their forms, as there is a reliable relationship between form and meaning being adopted, even though this is considered to be incongruent. This may benefit the learning of categorical distinctions in the language when the vocabulary size is small, as the systematicity is working to promote the learning of categories. In the present experiment, where there is no incongruent condition, the sound-symbolism has a clearly observable facilitative effect for learning of categories in the language.

### 3.3 Experiment 3: Learning from a fully arbitrary language

So far we have presented two experiments which look at how vocabulary size influences learning of a mixed vocabulary comprised of mappings which were either sound-symbolically congruent or incongruent ([Chapter 2](#)), in addition to a fully sound-symbolically congruent language (this Chapter). Yet there is still no direct evidence of how a fully arbitrary language would influence learning. Thus by using a new vocabulary, where the relationship between form and meaning is random, we aim here to provide such evidence.

#### Method

**Participants.** Seventy-two participants took part in the experiment (45 female), with 24 assigned to each vocabulary size condition. This was selected to be the same sample size as in Experiment 1. Participants were undergraduate students from Lancaster University, with a mean age of 19.2 years ( $SD = 2.13$ , range 18-27). It was not required that participants spoke English as their first language (English first language speakers:  $n = 45$ ), but all participants spoke English competently. Informed consent was collected from each participant and ethical approval was obtained from Lancaster University's ethics committee.

**Materials.** The same set of 16 angular/rounded shapes that were used in [Experiments 1](#) and [2](#) were used as the visual stimuli for the present Experiment. For the auditory stimuli however, a new set of 16 non-words were created. This

was done in order to remove any possible relationship between sound and meaning, be it congruent or incongruent. To achieve this, we first generated a new inventory of auditory non-words. Thirty monosyllabic CVC non-words were recorded by the same native English speaker who recorded the auditory stimuli for [Chapter 2's](#) experiment. The non-words in this inventory were created using consonants from a set including plosives (/g/, /d/, /p/), continuants (/m/, /n/, /l/) and fricatives (/f/, /v/, /s/)<sup>8</sup>, with contrasting consonants being used for each word (i.e. a plosive would be used in onset position, but only a continuant or fricative would be used in coda position). Additionally, one of five vowels (/ʌ/, /ɛ/, /i/, /ɒ/, /a/) was used in the non-word, with a total of 30 non-words generated, all of which had no dominant phonological property associated with sound-symbolism (in contrast to the stimuli used in the previous experiment). To ensure that there were no differences between the acoustic properties of the recorded sounds, the properties of intensity, fundamental frequency (pitch), first, second and third formants were normalized using Praat ([Boersma & Weenink, 2015](#)), this was consistent with the properties of the auditory stimuli used by [Monaghan et al \(2012\)](#).

Data were then collected from a short questionnaire, where participants (n = 22) were presented with the auditory non-words over a pair of headphones, along with a 7-point Likert scale anchored by rounded and angular shapes (see [Figure 2.2 on p.38](#)), presented on a laptop computer. Participants would hear all

---

<sup>8</sup> Fricatives were chosen on the basis that they have previously been reported to be relatively neutral, with respect to their relationship to angular/rounded shapes, see [McCormick, Kim, List, and Nygaard \(2015\)](#).

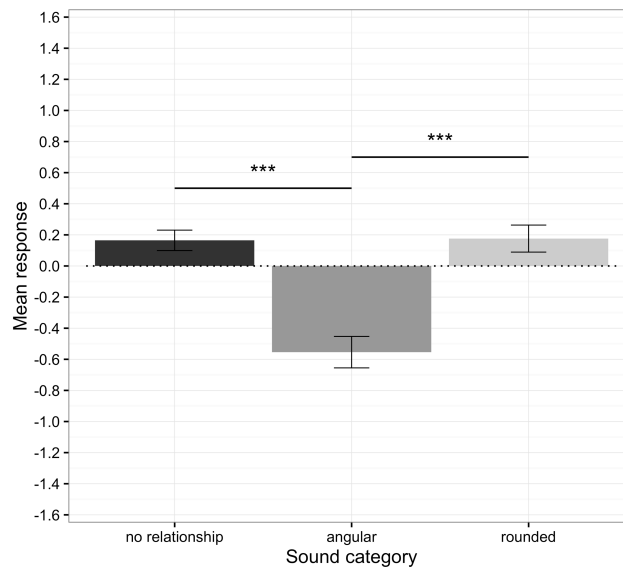
30 of the newly generated non-words, in addition to the 16 non-words presented during Experiment 1, allowing us to make a comparison between the two sets of stimuli. Each sound was presented twice throughout the questionnaire, with presentations randomized and the images anchoring the Likert scale counterbalanced (i.e. for half of the trials the rounded images were anchored on the left and then on the right for the other half). Participants were asked to rate each of the sounds based on how strong they felt it corresponded to either the rounded or angular shapes. This was done by selecting ‘0’ for no correspondence, ‘1’ for a weak correspondence, ‘2’ for a slightly strong correspondence or ‘3’ for a strong correspondence. Based on the mean ratings for the newly generated non-words, the 16 rated closest to 0 were selected and kept for further analyses (see [Table 3.1](#) for results).

To assess whether non-words were rated differently from each other, the 16 new non-words were compared to both the angular and rounded non-words used in [Experiment 1](#). We ran mixed-effects models with sound category (i.e. if the non-word was designed to be angular/rounded/no relationship) as a fixed effect, with questionnaire response as the dependent variable. The addition of sound category to the model significantly improved the fit ( $\chi^2(2) = 23.634$ ,  $p < .001$ ), with a significant difference between angular and rounded sounds (estimate = .73,  $t = 4.882$ ,  $p < .001$ ) and angular and no relationship sounds (estimate = .72,  $t = 5.550$ ,  $p < .001$ ), however there was no significant difference between the rounded and no relationship sounds (estimate = .01,  $t = .088$ ,  $p = .931$ ), see [Figure 3.3](#) for results. This final set of 16 non-words were then



**Table 3.1.** List of phonetically transcribed non-words used during Experiment 2, with the mean response rating from the questionnaire analysis (positive values represent preference for rounded shapes, negative values represent

Non-word	Mean rating (SD)
/gʌs/	-0.27 (1.58)
/fɒd/	-0.16 (1.53)
/mag/	-0.14 (1.53)
/vʌd/	-0.05 (1.49)
/nɛf/	0.02 (1.55)
/vin/	0.14 (1.33)
/nad/	0.14 (1.35)
/vɒg/	0.14 (1.49)
/dɒf/	0.25 (1.30)
/dam/	0.3 (1.18)
/gɒv/	0.34 (1.16)
/niv/	0.34 (1.45)
/fɛn/	0.36 (1.35)
/mɛv/	0.39 (1.27)
/pʌv/	0.41 (1.43)
/gal/	0.43 (1.42)



**Figure 3.3.** Mean accuracy of responses for presentation of auditory stimuli designed to incorporate either rounded or angular phonetic characteristics (stimuli from Chapter 2) and when there is no intended relationship (stimuli from experiment 2, this Chapter). Positive values represent preference for rounded shapes, negative values represent preference for angular shapes. Error bars show SEM. \*\*\*  $p < .001$ .

(estimate = .73,  $t = 4.882$ ,  $p < .001$ ) and angular and no relationship sounds (estimate = .72,  $t = 5.550$ ,  $p < .001$ ), however there was no significant difference between the rounded and no relationship sounds (estimate = .01,  $t = .088$ ,  $p = .931$ ), see Figure 3.3 for results. This final set of 16 non-words were then mapped randomly to one of the 16 rounded or angular shapes. The full set of words can be found in Table 3.1.

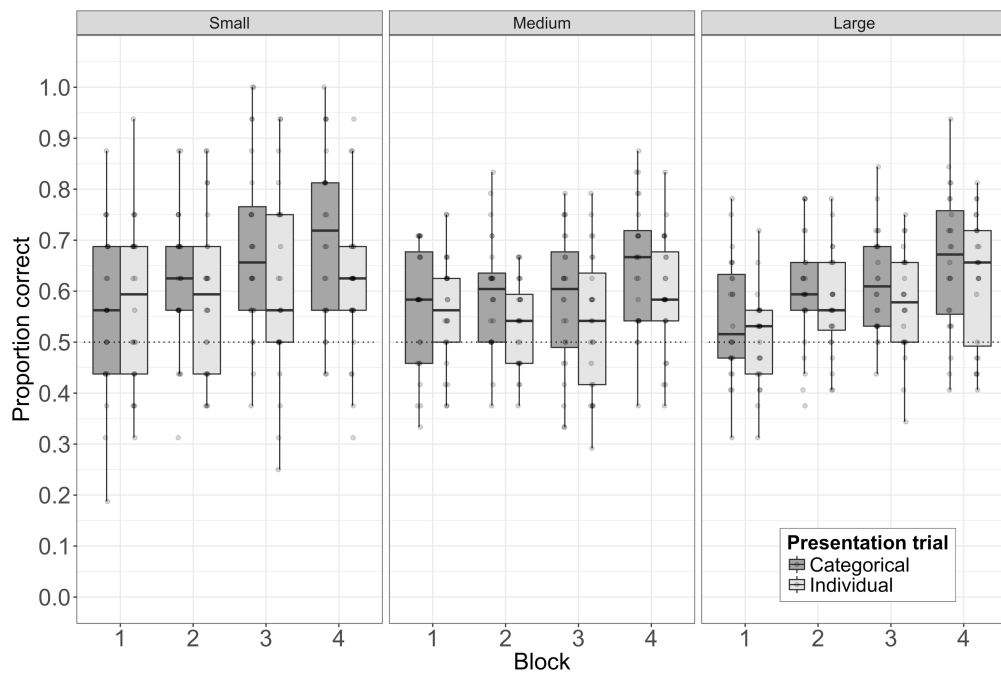
**Procedure.** The procedure was identical to Experiment 1.

## Results and discussion

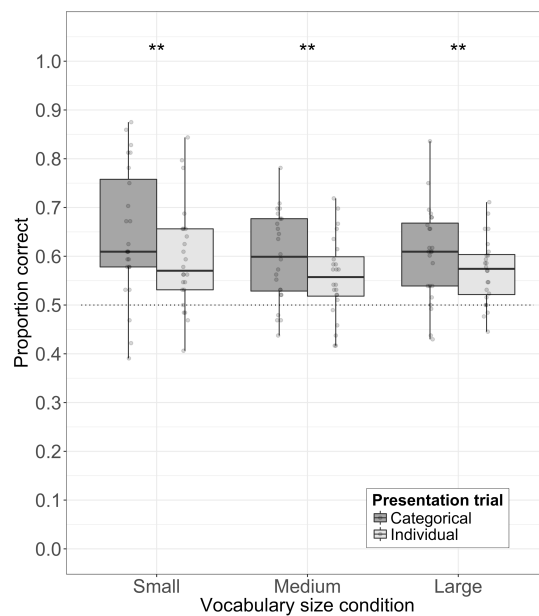
Following the same analysis as used in Experiment 1, a sequence of generalized linear mixed-effects models were fitted to our data to predict response accuracy.

The inclusion of block significantly improved model fit ( $\chi^2(1) = 58.336, p < .001$ ), indicating that over the course of the experiment participants were improving the accuracy of their responses. Further analyses using one-sample t-tests showed that for all experimental conditions, performance was significantly above the 50% chance level at the last block of the experiment (see [Figure 3.4](#) for results). The inclusion of vocabulary size did not significantly improve model fit ( $\chi^2(2) = 3.917, p = .141$ ), crucially indicating that there were no significant differences in overall accuracy across vocabulary size conditions. The inclusion of presentation type (categorical/individual) significantly improved model fit ( $\chi^2(1) = 23.520, p < .001$ ), indicating that accuracy on categorical trials was higher than individual learning trials. This was the case for each of the three vocabulary size conditions, with small, medium and large conditions showing significantly higher accuracy for categorical trials (small:  $\chi^2(1) = 7.740, p = .005$ ; medium:  $\chi^2(1) = 8.572, p = .003$ ; and large:  $\chi^2(1) = 8.167, p = .004$ ), see [Figure 3.5](#) for results). However, we did not find a significant improvement to model fit when the interaction term vocabulary size x presentation type was added ( $\chi^2(4) = 4.520, p = .340$ ), indicating that as vocabulary size varied, the differences in accuracy for categorical and individual learning trials remained.

The addition of a three-way interaction term between block x vocabulary size x presentation type did significantly improve the model fit ( $\chi^2(9) = 18.282, p = .032$ ), indicating that there was a difference between accuracy for categorical and individual learning trials over the course of the experiment, depending on the size of the vocabulary. In order to understand this interaction, we tested further



**Figure 3.4.** . Proportion of correct responses by block, for same and different category presentations, by vocabulary size condition) for Experiment 3. Dots represent individual subject data. Dotted line shows 50% chance level.



**Figure 3.5.** Proportion of correct responses in different category presentation trials (categorical learning) and same category presentation trials (individual word learning) for Experiment 3. Dots represent individual subject data.  $**p < .01$ .

models that examined performance during each vocabulary size condition separately, allowing us to explore the two-way interaction of block and presentation type (categorical/individual word learning). For the small vocabulary size, there was a significant improvement to model fit when the interaction term was included ( $\chi^2(1) = 7.1.05, p = .008$ ), with follow up analyses revealing accuracy on categorical trials improving over the course of the experiment ( $\chi^2(1) = 22.426, p < .001$ ), but not for individual word learning trials ( $\chi^2(1) = .985, p = .321$ ). This contrasts to the medium and large vocabulary size conditions, where the addition of the interaction term did not significantly improve model fit (medium:  $\chi^2(1) = 1.229, p = .268$  and large:  $\chi^2(1) = .003, p = .955$ ), indicating that accuracy for both categorical and individual word learning trials improved over the course of the experiment. See [Figure 3.4](#). This indicates that over the course of the experiment there was an improvement in accuracy in every condition apart from when the vocabulary size was small and the trials tested learning of individuated meanings in the language.

### 3.4 Across experiments comparison

If we consider that Experiments 2 and 3 assessed how artificial languages are learnt when the mappings are either sound-symbolic or have no direct relationship between form and meaning, alongside the experiment conducted in [Chapter 2](#), where congruent and incongruent mappings were presented (as part of one artificial language), then it would be of interest to compare the results of the fully sound-symbolic and no relationship languages together. This would provide

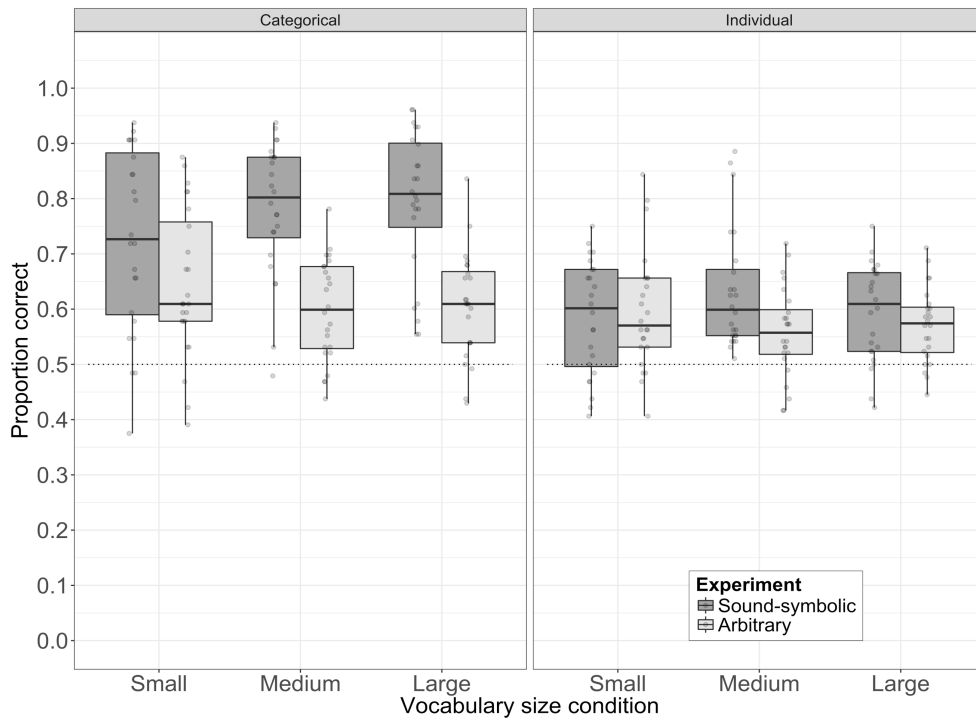
an additional analysis of how each experimental condition influences learning performance, which will draw the experiments together. In order to do this each of the experiments will be compared using generalized linear mixed-effects modelling, with fixed effects of block, vocabulary size, same/different category presentation type and experiment added sequentially to the model, allowing for a three-way interaction, between vocabulary size x presentation type x experimental condition to be investigated.

Additional comparisons using the experiments from this Chapter and from [Chapter 2](#) have also been conducted, whilst the focus of this chapter is to examine how fully sound-symbolic or arbitrary languages may be affected by vocabulary size, these additional analyses can be found in the [Appendix](#), with a full analysis of how Experiment 1 compared to Experiments 2 and 3.

### **Fully congruent and no relationship**

The addition of experiment as a fixed effect revealed a significant improvement to model fit ( $\chi^2(2) = 30.66, p < .001$ ), with accuracy in the fully congruent condition being significantly greater than the no relationship. Furthermore, the addition of the three-way interaction between vocabulary size x presentation type x experimental condition also significantly improved model fit ( $\chi^2(8) = 22.16, p = .005$ ).

To understand this interaction, we then carried out further analyses on the separate presentation type trials (categorical or individual learning) and the different vocabulary sizes. For the categorical trials, the inclusion of experimental



**Figure 3.6.** Proportion of correct responses in different category presentation trials (categorical learning) and same category presentation trials (individual word learning), comparing Experiment 2 (sound-symbolic) and Experiment 3 (arbitrary). Dots represent individual subject data.

condition significantly improved model fit for all vocabulary sizes (small:  $\chi^2(1) = 4.13, p = .042$ ; medium:  $\chi^2(1) = 26.85, p < .001$ ; large:  $\chi^2(1) = 24.35, p < .001$ ), indicating that accuracy was higher for the fully congruent condition, when compared to the no relationship condition. See Figure 3.6 for results.

For the individual learning trials, the inclusion of experimental condition did not significantly improve model fit for the small or large vocabulary size (both  $p$ 's  $> .05$ ), indicating that there was no difference in accuracy between the fully congruent or no relationship conditions. However, there was a significant improvement to model fit for the medium vocabulary size ( $\chi^2(1) = 6.75, p =$

.009), indicating that accuracy was higher for the fully congruent condition, when compared to the no relationship condition. See [Figure 3.6](#) for results.

These results demonstrate that there is a clear benefit of sound-symbolism over mappings with no relationship when learning categorical distinctions, and that this is observed in all of the vocabulary size conditions. However, there is no evidence to suggest that a fully sound-symbolic language benefits the learning of individual meanings in a small vocabulary size, however the medium size vocabulary did show evidence of this advantage, but these results are thus far difficult to explain and may be an anomalous result. Additionally, there is no evidence to suggest that a language with no relationship between form and meaning is more beneficial in a larger vocabulary size.

### 3.5 General Discussion

Within these two experiments, we have examined the influence of vocabulary size when a language is made up of only sound-symbolic or arbitrary form-meaning mappings. In Experiment 2, the results demonstrate that there was a general trend for greater accuracy during trials testing the learning of broad categories, when compared to trials that tested learning of individuated meanings, consistent across all vocabulary sizes. This indicates that sound-symbolism promotes the learning of categorical distinctions even when the vocabulary size is small, but particularly so when it is large. This is consistent with previous reports of sound-symbolism in the vocabulary, specifically systematicity, where it has been demonstrated that having a non-arbitrary relationship which identifies the



category of the word, for instance through the use of specific phonological properties, is not only advantageous, but is also present in natural language ([Mongahan et al., 2012](#)).

In Experiment 3, we demonstrated that within a fully arbitrary language, a learner does not incur a cost for learning individuated meanings, even when the vocabulary size is increased. Interestingly, it appears that individual learning actually improves over time when the vocabulary size is large, a pattern that is not observed in the small vocabulary size, this would suggest that as the vocabulary size grows, and new words are added to it, then an arbitrary system would be most advantageous, a finding that is supported by previous computational models ([Gasser, 2004](#); [Monaghan et al., 2011](#)).

# Chapter 4

## Predictors of lexical stability in an artificial language

---

### 4.1 Introduction

Languages undergo change over time. Be it the lexical, grammatical or phonological properties that language users adopt, there are pressures from social and economic factors which contribute to this change ([Labov, 2001](#)), as well as pressures from human cognition that also drive change ([Pagel et al., 2007](#)).

Indeed, such modification over time has resulted in the striking diversity and variety of languages being used today, with estimates of over 7,000 documented living languages today ([Lewis, Simons, & Fennig, 2016](#)) each of which is somehow distinguishable from every other language<sup>9</sup>. However, this is not to

---

<sup>9</sup> Whilst this is a very simplified notion of what constitutes a distinct ‘language’, it serves to highlight the large variation present in the world’s languages. For a more formal framework and discussion on defining and classifying the world’s languages, see [Cysouw and Good \(2013\)](#).

claim that languages share no common ancestral history. In fact, phylogenetic analyses (Gray & Jordan, 2000; Gray & Atkinson, 2003), have highlighted in detail that languages, like living organisms, evolve and diverge by descent, with innovations and modifications being introduced over time. Coupling this with evidence of diachronic change in language (Bybee, 2001), where shorter term processing constraints act to modify the language, one can begin to establish a solid explanation of linguistic variation and diversification.

Evidence of linguistic change across languages can be demonstrated neatly by contrasting the word forms used by two languages to represent a given meaning. For example, the forms used in English and French to represent the number *fourteen* are /fɔːti:n/ and /ka.tɔʁz/ (*quatorze*) respectively. Despite these languages descending from the Indo-European language family, the difference in word form highlights that over time, these two languages came to adopt different lexical conventions into their vocabularies for the same meaning (see Calude and Verkerk, 2016, for more discussion on variation in numerals).

However, these differences between languages can be much more subtle, it can be observed that certain items within a vocabulary appear to be much more resilient to change, resulting in certain words being relatively more conserved across different languages. This can again be demonstrated by comparing the English word forms used to represent the meaning of the numeral *two*, where in English the form *two* (/tuː/) is used, whilst in French it *deux* (/dø/). This example can even be extended to languages that are phylogenetically more distant to each other, such as in Greek, where the form used is *δύο* (/dýo/). Thus, language

change may operate on two distinct levels, where a word is either replaced by a completely unrelated form or adjusted slightly over time.

Understanding why certain properties of language undergo more dramatic and rapid change, whilst others appear to conserve their states and are thus more resistant to modification, will be the core topic of investigation in this chapter, presenting a cognitive explanation for what drives such changes in the vocabulary. Through a series of behavioural experiments, quantitative evidence will be generated that demonstrates how measures of cognitive influence can be used to predict differences in the types of change word forms undergo. This will build on previous theoretical and empirical findings, but achieved through the use of a novel and controlled laboratory paradigm.

Given that previous findings have used data from comparative studies of natural language to generate evidence for cognitive theories of language change, where many different variables need to be accounted for during analysis (see [Monaghan, 2014](#)), the approach taken here will ensure that there is considerable control and isolation of each of the variables of interest. In doing so, the aim will be to use psycholinguistic properties of words to predict when either large, small or no changes are likely to occur.

By adopting an artificial language learning paradigm, psycholinguistic properties can be manipulated during a learning phase, where the participant is exposed to, and expected to learn, a language. Then, once the learning phase has been completed, the participant's ability to reproduce the language can be measured during a testing phase. In addition to this, the data will then apply an

innovative statistical approach, whereby different types of linguistic change can be identified. This will aim to make quantitatively important distinctions between whether word forms undergo lexical replacement, where a participant has produced a word during testing that dramatically differs from the word they were exposed to during training, or alternatively, lexical adjustment, where the word produced during training is incorrect, but still resembles the training word. Thus, we present here an approach that will provide a deeper understanding of the way cognition may reflect processes of language change in the vocabulary.

## 4.2 Effects of frequency on language change

An undeniable feature of the vocabulary of any language is that some words will occur often, whilst others occur comparatively less so ([Zipf, 1936](#); [Piantadosi, 2014](#)). The number of times we might produce the word ‘*the*’, dramatically differs from the number of times the word ‘*mizzenmast*’ is likely to be used<sup>10</sup>. Early research interested in exploring this phenomena, highlighted that there was a strong awareness of the variation in word use frequencies. Studies that assess how accurately the frequencies of words and letters in natural language can be predicted, demonstrated that estimates were produced with relatively high levels of accuracy ([Attneave, 1953](#); [Balota, Pilotti, & Cortese, 2001](#); [Shapiro, 1969](#)).

Whilst this reveals that we may be aware of whether a word is high or low in frequency, there has also been extensive research on the effects word

---

<sup>10</sup> examples from the most and least frequent words from the SUBTLEX-UK database ([van Heuven, et al., 2014](#)).

frequency has for language processing (see [Ellis, 2002](#) for review). Within this vast literature, there has been robust evidence that highlights the processing advantages resulting from having reliable statistical regularities embedded within a language. These effects are wide ranging and can be observed in areas critical to using language effectively, such as in processes of language acquisition ([Monaghan, Chater, & Christiansen, 2005](#); [Saffran, Aslin, & Newport, 1996](#); [Smith & Yu, 2008](#)), comprehension of grammatical constructions ([Ford, Bresan, & Kaplan, 1982](#), [Juliano & Tanenhaus 1993](#); [Jurafsky, 1996](#)) and language production ([Bybee, 2001](#); [Bybee & Scheibman, 1999](#), [Gahl & Garnsey, 2004](#)), see also [Diessel \(2007\)](#) for review.

If there is a bias towards more reliable and accurate acquisition, comprehension and production of high frequency occurrences in language, then this would have implications for the way that features of language may experience change over time. It could be argued that this processing advantage can dramatically increase the chance of frequently occurring items resisting the processes of language change, whilst consequently leaving items that occur much less frequently highly susceptible to these processes. Evidence for this can be observed in both diachronic linguistics, where proto-words are reconstructed and then studied to investigate change, and also in phylogenetic approaches, where changes in cognates are examined.

First, diachronic analysis has uncovered two apparently contradictory effects of frequency. As outlined by [Bybee and Thompson \(1997\)](#), high frequency tokens can drive language change in a reductive manner, whilst also

generating a conservational force. To demonstrate how reduction can occur dominantly in highly frequent occurrences, [Bybee and Thompson](#) highlight the loss of the final phonemes /t/ and /d/ when used in spoken in high frequency words such as ‘*went*’, ‘*just*’ and ‘*and*’, whilst this reduction does not occur in lower frequency words (see also [Bybee, 2002](#); [Hay et al., 2015](#); [Hooper, 1976](#); [Scheibman, 2000](#)).

The reason why reduction of high frequency tokens occurs, is argued to be as a result of neuromotor routines ([Boyland, 1996](#); [Bybee, 2002, 2006](#)). Given that higher frequency tokens are used with greater repetition than those that have lower frequency of use, this increased repetition allows the tokens to become more fluent and automatized. This leads to reduction and what has been described as ‘chunking’ ([Christiansen & Chater, 2015](#); [Ellis, 1996](#)), which is the merging of complex units into a more efficient and compressed singular unit. Thus, it is through this process that high frequency tokens can be subjected to small changes, but still remain easily processed in language use. Such a view is in line with [Zipf’s \(1949\) principle of least effort](#), whereby highly frequent words or utterances are optimised for efficient communication. If production can be made more efficient without incurring relative processing costs, then a pressure for efficiency will be introduced on to the language user which will ensure production effort is minimised. Such a pressure will only be introduced for high frequency utterances however, as low frequency utterances must prioritise accurate production, given the relative likelihood of inaccurate processing by the listener.

On the other side of diachronic language change, there is also evidence to suggest what at first may appear to contradict the idea that high frequency of use is the focus of change. Instead, much research has shown that high frequency of use actually enables those tokens to be conserved in the language, resulting in a frequency effect that leaves low frequency tokens more susceptible to change, whilst high frequency tokens are resistant to change. Similar to the process of reduction, conservation works against the general rules adhered to by the majority of language use.

A prime example of conservation can be seen in patterns of regularisation, as shown in [Lieberman et al's \(2007\)](#) examination of verb regularisation over the course of Old, Middle and Modern English; throughout which time the regular past tense suffix *-ed* was adopted<sup>11</sup>. Despite this rule emerging and consequently changing the past participle forms of many verbs, there still remains a number of verbs (less than 3%) that don't adhere to this convention, which notably are the most frequently used verbs in English. This is evidence therefore of a frequency effect that allows the most frequently used words to remain in a stable state, whilst other less frequent words are forced to conform to a pattern, even if that means dramatic replacement through analogical change.

Again it is argued by [Bybee \(2001, 2006\)](#), that the persistent repetition of these high frequency tokens in the language is what allows them to be conserved in this way. Despite irregular constructions being used in the language, they are

---

<sup>11</sup> Also note that there are other examples of conservation from regularisation as demonstrated by [Krug \(2003\)](#) and also by [Bybee \(2006\)](#), where evidence of this process in words other than verbs is produced.



able to prevail against the forces of both time and conformity. This entrenchment of high frequency tokens allows for their strong representation in memory and ease of acquisition, which would therefore enable them to withstand any possible shifts away from their current state (Ellis, 2002; Langacker, 1987).

Interestingly, these two frequency effects of reduction and conservation work alongside each other effectively, where high frequency of use allows for small local adjustments to be introduced that enable greater fluency of processing. The impact of which, is hypothesized to strengthen the mental representations of the tokens and makes them more resistant to any universal changes that might be taking place in the language.

Whilst diachronic studies provide insights into language change on a local and perhaps temporally specific scale, the alternative of phylogenetic inference allows for a much wider scope of language history to be explored. Through this approach, claims about language change across different language families and within much deeper measurements of time can not only be explored, but explored quantitatively (Mace & Holden, 2005, Pagel, 2009). More precisely they can offer support for the effect of frequency on language change from an evolutionary viewpoint.

If it is to be claimed however, that frequency has a direct effect on the way a language evolves, then it is important to assess whether measurements of frequency are stable across the world's languages, else generalisations made about one language may not be applicable to other languages. Evidence in support of a shared pattern of word use frequency has been shown by Calude and

[Pagel \(2011\)](#), who demonstrated that across six different language families, frequency of use for 200 fundamental meanings was highly inter-correlated. This finding gives support to claims about how frequency affects language change, not simply in one language, but across the world's languages.

A core finding from phylogenetic analysis for an effect of frequency in language change has been reported by [Pagel et al \(2007\)](#). Here, it was shown that the rate at which a word is replaced by a new, unrelated form, can be accurately predicted by the number of times that word is used in everyday contexts. [Pagel et al](#) used items from the Swadesh list of fundamental vocabulary meanings ([Swadesh, 1952](#)), these items are considered to be essential to all human languages, therefore these items would be suitable for large time scale analysis, such as that used in phylogenetic studies of language change. The results showed that frequency of use was a reliable predictor for rates of replacement. Similar to the effect of conservation as reported by [Bybee and Thompson \(1997\)](#), higher frequency words were shown to have greater lexical stability, in comparison to words with lower frequency of use, which undergo replacement at a faster rate. This finding demonstrates that lower frequency words will adopt unrelated forms more rapidly than words exhibiting high frequency of use, thereby conserving their cognate form for much longer.

In order to test if these previously reported frequency effects can be observed and explored experimentally, here we test through the use of an artificial language learning paradigm, how exposure to words that vary in frequency will directly affect how well the words are learnt and reproduced.

Further to this, the paradigm used will enable an assessment of what types of error occur when words are not accurately recalled, highlighting important distinctions between effects of replacement/conservation and more subtle changes of adjustment, where the words may only change slightly, but do so in a way that could offer advantages similar to those of reduction. Therefore, this offers a novel approach to studying how frequency effects on language change can be explored in the laboratory.

We predict that when a word is high in frequency, there will be significantly robust processing, meaning that the word will undergo very little change during recall. In contrast, low frequency words will not have such processing advantages, leading to much more replacements being introduced, creating a dynamic in the language where frequency influences where large and small changes may occur.

## **Experiment 4: Lexical Stability and Frequency of Words**

### **Method**

**Participants.** The experiment was completed by 21 undergraduate students from Lancaster University (14 female) with a mean age of 18.95 years ( $SD = 0.89$ , range = 18-21). All participants were proficient in English and received course credit for participating.

**Materials.** Participants were presented with an artificially created language, consisting of written non-word strings (representing the forms in the language), each of which was mapped onto an abstract image (representing the

meanings in the language). The meanings were from the same set used by Kirby et al (2015), which comprised 12 images that varied in shape (4 possible shapes) and texture fill (3 possible fills), each image also had its own unique appendage. See Table 4.1 for the full set of meanings used.

There were 12 written word forms that were generated by combining consonants and vowels together to make a 6 letter string with a CVCVCV structure. The strings were derived from an inventory of 8 consonants (g, h, k, l, m, n, p, w) and 5 vowels (a, e, i, o, u). To minimize any underlying structure in the language, there were no duplicated CV syllables within any individual word, whilst also ensuring that no consonant or vowel would occur more than once. Additionally, no consonant would occur more than two times in each of the consonant positions across the set of words, and vowels would occur no more than three times in each vowel position (i.e. for all 12 word forms, there would only be a maximum of two that started with any given consonant). Word forms were then inspected to see if they resembled any English words, if this was the case the form would be replaced with an alternative. A Mantel test was performed on the language set to ensure that no systematic structure was present (see Analysis section for more details on this test). Using this same process, 4 different language sets were generated and each participant was randomly assigned a language to learn from during the experiment. An example language can be found in Table 4.1.













To investigate the effect of frequency on participants' learning, we assigned four mappings to either a low, medium or high frequency condition.

High frequency mappings occurred 6 times more often than the low frequency ones, and medium frequency mappings 3 times more often than the low frequency ones.

**Procedure.** Participants completed a training phase during the experiment where they were given a number of exposures to the form-meaning mappings, after they completed the training they were then asked to complete a testing phase. Before beginning the experiment, participants were told that they would be learning an alien language, where they would see an image and a written word that is associated with the image and that their task was to learn the associations. The experiment was run using E-Prime 2.0 and was completed on a computer. During the training phase, all participants were given 120 exposures, over the course of 3 blocks. In the low frequency condition, each mapping was presented once per block, for the medium frequency condition, mappings were presented 3 times per block, for the high frequency condition, mappings were presented 6 times per block. During each presentation, participants were shown just the meaning for 1 second, followed by the meaning with the corresponding written form underneath for 5 seconds. After each exposure, participants were prompted with just the meaning again and asked to type the corresponding written form, this was to ensure effective and accurate learning during training.

After completing the training, participants then completed a testing phase. During this phase each of the meanings was shown to the participant once. During each presentation, they were asked to type a response that they believed was the associated form for that meaning. Before beginning the experiment,

**Table 4.1.** Example language sets used in Experiments 4, 5 and 6, showing the meaning, form mapped onto the meaning and the frequency/length/acquisition condition the form-meaning mapping was allocated to.

	Experiment 4		Experiment 5		Experiment 6	
Meaning	Form	Frequency	Form	Length	Form	Acquisition
	punima	medium	punima	medium	punima	late
	howuna	low	wapohe	medium	howuna	early
	lomapi	low	gilo	short	lomapi	early
	mehapu	high	howu	short	mehapu	late
	wapohe	medium	kuga	short	wapohe	late
	kugawe	low	henumiku	long	kugawe	early
	gapulo	high	loma	short	gapulo	late
	muhego	medium	gapulona	long	muhego	early
	giloku	low	nihewapi	long	giloku	early
	henumi	high	mekapuwe	long	henumi	late
	nikewa	high	muhego	medium	nikewa	late
	nakilu	medium	nakilu	medium	nakilu	early

participants were instructed to not leave the response blank and to always try and give a response that could be part of the language. All presentations in training and testing were shown in a randomized order.

## Results

**Learnability.** To assess the effect of frequency on participants' learning, measures of performance were calculated based on the training language they received and the testing output they produced. Firstly, accuracy was measured by directly comparing the training forms to those generated during testing, if both forms were identical then the participant would receive a score of 1, if they did not match they received 0. Thus, the maximum total score a participant could receive was 12.

Secondly, we calculated the Levenshtein edit distance ([Levenshtein, 1966](#)), this measurement compared each individual character in the training and testing forms to one another, providing a more precise measurement of learning error than accuracy. If a character produced during testing was changed, inserted or deleted from that of the original training form, then a participant would receive 1, if the characters matched then they would receive 0. A total sum was then calculated by comparing all the characters in the training and testing forms (this was calculated using the stringdist package ([van der Loo, 2014](#)) in R ([R Core Team, 2011](#))). Following [Kirby et al's \(2008\)](#) analysis, a normalised Levenshtein distance was also calculated by taking the summed Levenshtein distance, then dividing it by the length of the longest of the training-testing forms. This

additional measure was used in order to take into account the possibility that some words may have been recalled in a way that shortened or lengthened the word.

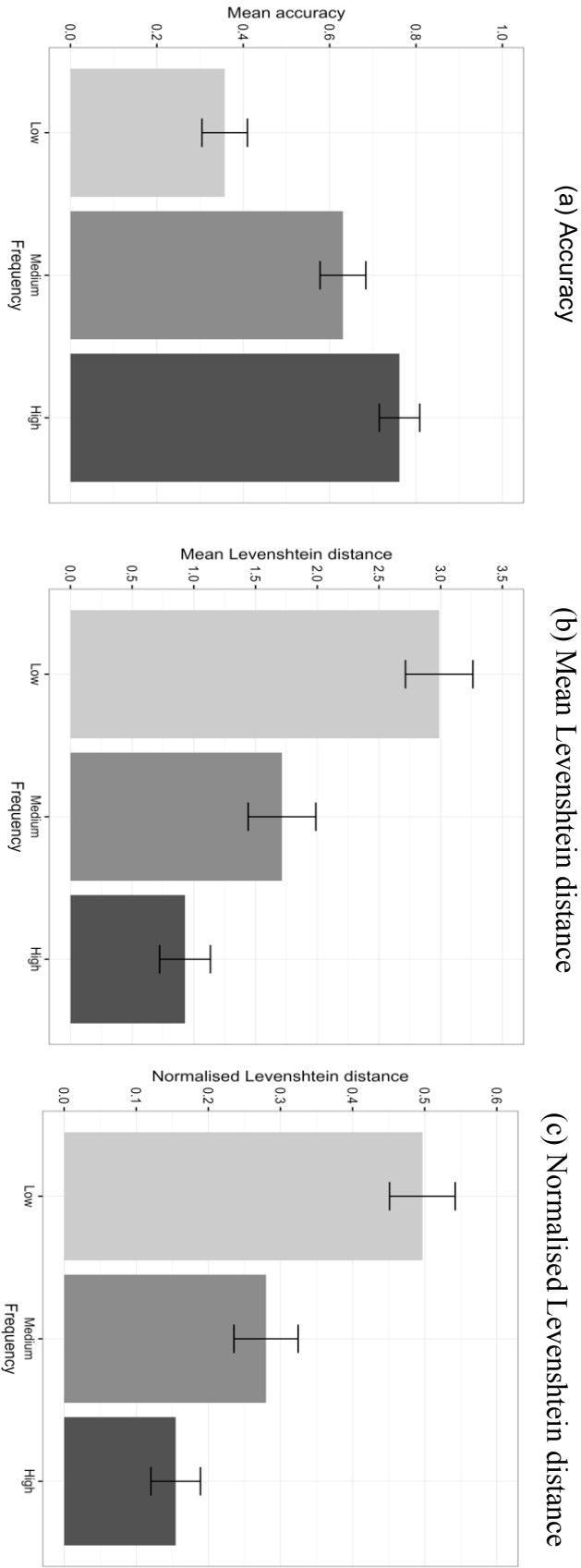
In the analysis conducted on the data, we modelled the probability (log odds) of our three dependent variables (accuracy, Levenshtein distance and normalised Levenshtein distance), accounting for the variation across participants and language sets. We performed a sequence of Linear Mixed-effects Models<sup>12</sup> (Baayen, 2008; Jaeger, 2008), specifying first the random effects, then adding in the fixed effect of frequency, which was treated as an ordinal factor with levels ordered from low < medium < high. After adding the fixed effect to the model, we ran likelihood ratio test comparisons, comparing the new model to the previous one with only random effects. This showed whether the inclusion of the new term significantly improved the fit of the model.

For the measurement of accuracy, frequency significantly improved model fit ( $\chi^2(2) = 41.014, p < .001$ ), indicating that accuracy increased as frequency increased (estimate = 1.78,  $SE = .31, z = 5.66, p < .001$ ). Additionally, the Levenshtein distance measure also showed a significant improvement to model fit ( $\chi^2(2) = 45.501, p < .001$ ). This was also the case for the normalised Levenshtein distance measure where the model fit improved ( $\chi^2(2) = 46.24, p < .001$ ), indicating that error decreased as frequency increased (Levenshtein

---

<sup>12</sup> For the analysis on the accuracy measurement, Generalized Linear Mixed-effects models were performed.





**Figure 4.1.** The effects of frequency condition on learning. (a) shows mean accuracy of participants' recall where lower scores represent lower accuracy, (b) shows mean Levenshtein distance where lower scores represent higher accuracy, (c) shows normalised Levenshtein distance where lower scores represent higher accuracy. Error bars show standard error of the mean by participants (SEM).

estimate = -1.46,  $SE = .21$ ,  $t = -7.00$ , and normalised Levenshtein estimate = -0.24,  $SE = .03$ ,  $t = -7.04$ ). See [Figure 4.1](#) for results.

**Rates of replacement and adjustment.** Next, our analyses aimed to investigate whether frequency could be used to predict rates of replacement and adjustment in the participant's testing output. Although our previous analysis quantified errors made during testing (similar to standard psycholinguistic experiments on word learning), our next analyses focused on distinguishing between the types of change that have resulted from recall, providing a means to explore hypotheses about processes of linguistic change. In order to do this, we first had to determine a criterion to distinguish responses as either replacements or adjustments; whereby we could quantifiably establish whether a participant had produced a word that either resembles (adjustment) or differs (replacement) from the target word.

This was achieved by comparing each individual Levenshtein edit distance to the average Levenshtein distance of a word generated at random to the words that were used to train the participants. To achieve this we computed a Monte Carlo test of sampling words randomly run 1,000,000 times. For each sample, words were composed of the 8 consonants and 5 vowels that were used in the initial starting languages, in addition to conforming to the same 6 letter length and CVCVCV structure.

We then calculated the mean Levenshtein distance between each of the words from the initial training languages and the words from the Monte Carlo sample. Over the 1,000,000 sampling runs, this produced an average edit distance

$$CT = \frac{\mu_{LD} - \alpha\sigma}{len}$$

**Figure 4.2.** Calculation of critical threshold (*CT*) for adjustment/replacement classification. Where  $\mu_{LD}$  is the mean Levenshtein distance between the training word and Monte Carlo sample of 1,000,000 permutations,  $\alpha$  is the value at which a Z-score is significant at the 95% confidence level here 1.64 is used,  $\sigma$  is the standard deviation of  $\mu_{LD}$ , *len* is the length of the input word.

of 5.005 (SD = 0.891) between all pairs of words, which represents the mean distance between a word generated at random and the original version of that word. Next we converted these values to determine the critical distance at which a particular production is more likely than chance to be distinct (corresponding to a Z-score distance of -1.64,  $p < .05$ ). We used this cut-off to determine whether a participant's Levenshtein distance was significantly closer to an unrelated word, or its original training word form. This produced our critical threshold value of 3.54. For the normalized Levenshtein distance, this value was 0.590 (3.54/6, the length of the words), see [Figure 4.2](#) for equation.

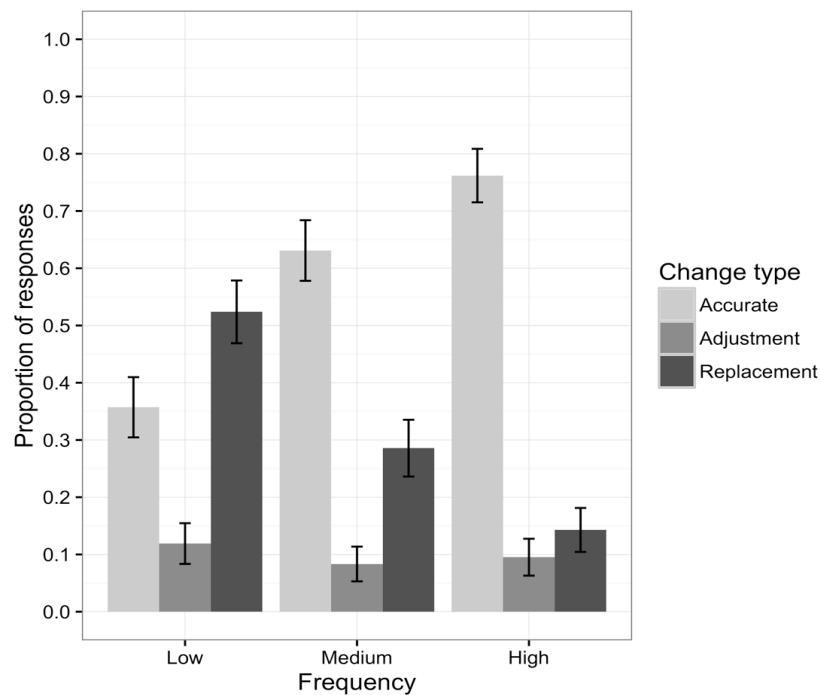
Any response that was incorrect and had a Levenshtein distance that was greater than the critical threshold value was coded as a replacement, any response that was incorrect and smaller than the threshold value was coded as an adjustment, whilst any response with a Levenshtein distance of 0 would be coded as accurate. For example, if a training word was *giloku* and during testing the recall was *gilowe* then this would be coded as an adjustment (Levenshtein distance = 2), whereas if the recall was *lukehe*, then this would be coded as a replacement (Levenshtein distance = 6).

**Table 4.2.** Contingency table showing counts of different response types by frequency condition.

	Response type		
	Accurate	Adjustment	Replacement
Low	30	10	44
Medium	53	7	24
High	64	8	12

We performed a sequence of generalized linear mixed-effects models on our data to see if the fixed effect of frequency could reliably predict rates of replacement and adjustment, with random effects of subject and training language set, constructing separate models for the two dependent variables of replacement and adjustment. For rates of replacement, frequency significantly improved model fit ( $\chi^2(2) = 42.249, p < .001$ ), indicating that as frequency decreases, rates of replacement increases. For rates of adjustment, frequency did not significantly improve model fit ( $\chi^2(2) = 0.632, p = 0.729$ ), indicating that there was no difference between the rates of adjustment and the frequency of the word being presented during training.

Next, our analysis aimed to see if there was a difference between the number of responses coded as accurate, adjustment or replacement. To achieve this, we first constructed a contingency table of counts for each of the response types by frequency condition, see [Table 4.2](#). Next, we performed a log linear analysis on this data to see if there was an interaction between response type and



**Figure 4.3.** Proportion of testing responses classified as either accurate, adjustments or replacements in Experiment 4.

frequency. A Likelihood Ratio Test (LRT) revealed that the interaction term was significant ( $\chi^2(4) = 33.464, p < .001$ ). This indicated that as frequency varied, the number of replacement and accurate responses significantly differed (estimate = 1.45,  $SE = .28, z = 5.224, p < .001$ ) and a marginal difference between replacements and adjustments (estimate =  $-.69, SE = .37, z = 1.87, p = .061$ ).

To understand this, the data were then subsetting by frequency condition and further log-linear analyses were run. For low frequency words, there were no significantly more replacements than adjustments (estimate =  $-1.48, SE = .35, z = -4.229, p < .001$ ), but no significant difference between accurate and replacement responses was found (estimate =  $-.38, SE = .24, z = -1.618, p = .106$ ). For medium frequency words, there were significantly more replacements than

adjustments (estimate = -1.23,  $SE = .43$ ,  $z = -2.868$ ,  $p = .004$ ), in addition to significantly more accurate responses than replacements (estimate = 0.79,  $SE = .25$ ,  $z = 3.220$ ,  $p = .001$ ). For high frequency words, there was no significant difference between replacements and adjustments (estimate = -0.41,  $SE = .46$ ,  $z = -0.888$ ,  $p = .374$ ), but significantly more accurate responses than replacements (estimate = 1.67,  $SE = .31$ ,  $z = 5.321$ ,  $p < .001$ ). See [Figure 4.3](#).

## Discussion

In this experiment we aimed to test predictions that use frequency to explain why there is variation in the faithful production of lexical forms. The results showed that, in line with previous research, the frequency of a word significantly predicts the accuracy of production for that word. Hence, if a word exhibits high frequency of use within a language, then this will increase its chances of fidelity during use and therefore ensure that the form being used to represent its meaning has a greater likelihood of being conserved in the vocabulary. In contrast, words with a low frequency are more vulnerable to error, which consequently increases their potential to undergo change. Whilst this result is somewhat expected in light of the vast amount of work that highlights the relative robustness of frequency effects in learning (see [Diessel, 2007](#)), by implementing such an analysis in the present experiment, where the edit distance between the word presented during training and the word produced during testing is calculated, a finer grained measurement of error, rather than a binary judgment, can be calculated.

Moreover, by using such an error measurement, our results also uncover the different types of errors made by participants. Whilst a frequency effect on overall accuracy was expected to be observed, our results demonstrate in addition that low frequency words were more susceptible to much more dramatic errors, whereby the form of the word presented during training has been replaced by a completely different form during the production test. For example, an adjustment that was observed in one of the high frequency words was from the training form *lamupo* to the testing response of *lamupa*. Such a change had a Levenshtein distance of only 1, and the two forms retain a lot of similarity following the testing phase. In contrast, a replacement does not retain this similarity, for example a low frequency word that was observed to be a replacement in the language was *giloku* to *lukehe*. Here, a Levenshtein distance of 6 demonstrates that the original form bears no resemblance to the participants output, indicating that a completely new word form has been introduced to the language, which we attribute to the processing constraints that occur when words occur infrequently.

The provides evidence for the hypothesis that low frequency words are more vulnerable to errors that deviate dramatically from what has been observed during the learning process, whereas higher frequency words are less prone to such replacements. These results support [Pagel et al.'s \(2007\)](#) finding that frequency can be used to predict the rates of replacement within a language, with low frequency words being replaced at a more rapid rate than high frequency words.

Our analyses also investigated how frequency may also be used to predict rates of adjustment in a language. Whilst there were no significant differences in the rates of adjustment found in the present study, there may be plausible explanations for this result. If we reconsider the literature on linguistic change, there is strong evidence to suggest that when words are used more frequently, small changes are made to the word in order to maximize efficiency, such as reduction (for example see [Bybee and Thompson, 1997](#); [Zipf, 1949](#)). This type of adjustment would manifest as a result of the producer managing communicative effort, whereby any small changes in word form are likely to make articulation more efficient, such as final consonant deletion ([Jurafsky et al., 2001](#)). However, the artificial languages used in the present experiment conformed to a very strict CVCVCV structure, thus making any deviations in word length, through deletion for example, would be unlikely.

### 4.3 Effects of word length on language change

Another well documented effect on language use and processing is that found when there is variation in the length of a word, be it orthographic, phonological or syllabic length. Similar to the distribution of high and low frequency words in natural language, there is also an uneven division of labor between the number of small and long words in the vocabulary, with high frequency words being shorter than low frequency words ([Zipf, 1936, 1949](#); [Ferrer i Cancho & Solé, 2003](#)). This is evident if we return to the comparison of ‘*the*’, a short word with high frequency, with ‘*mizzenmast*’ a longer word, that is used comparatively less frequently.



Understanding why such a relationship between frequency of use and word length exists was initially addressed by Zipf (1949), with his *principle of least effort*. Here, it is claimed that the shortest word forms should be paired with the most frequently used items in a language, as this would produce a communicative system that is most efficient. For instance, if a language is to utilize the statistical distribution of word use with maximum efficiency, then it would be beneficial for those words to be shorter, as this would maximize the communicative efficiency of the lexical system. Having such a pressure to communicate efficiently, would mean that the forms used in a language would adapt to become shorter if they exhibited greater frequency of use within the population, thus allowing for quicker production for the speaker, in addition to greater ease of comprehension for the receiver.

In order to use a language with maximum efficiency however, the relationship between length and frequency may not represent the most effective system.

Developing from Zipf's work, Piantadosi, Tily, & Gibson (2011) demonstrated that word length can be more accurately predicted by the information content that the word contains. That is, if we take the view that languages are used in a manner that relies upon a variety of statistical dependencies, then language use and the probability that a word will be used in an utterance, is heavily reliant upon the context in which that word is being used in. That is, if a word is highly predictable in an utterance, then the less informative that word will be. If the word contributes very little to the utterance, then it would be more redundant, which in turn would mean it can be reduced in length, allowing for a more

optimal communicative utterance to be formed. Likewise, when a word contributes more information and is less predictable in the context of a given utterance, then that word should comprise more energy in the overall signal, meaning it should be longer.

This finding is complemented by work from [Aylett and Turk \(2004\)](#) and [Jaeger \(2010\)](#), who highlight the importance of distributing information across the course of an utterance in a uniform manner. Thus, in order for language to maximize communicative efficiency, an utterance must provide an optimal amount of information (i.e. neither too much or too little) in order to transmit the message accurately, without overburdening the production and comprehension systems involved. Incorporating variation in word length within language allows for this to be achieved. It is therefore argued that more redundant words, or ones that carry less information in the utterance, will be shorter in length and duration; whilst words that are less predictable, carrying more information content, will require longer length. This creates a trade-off between the effort involved in production of the utterance and ensuring accurate comprehension. With more predictable words carrying a greater chance of being understood, enabling them to undergo changes that might reduce production effort. This allows for less predictable words to take up more space in the utterance, increasing the chance of comprehension.

Indeed, there is copious evidence that shows the pressure to make communication in language more efficient, with speakers making adjustments to highly predictable words that reduce articulatory effort, but do not result in

comprehension errors. This is evident in the use of contractions in the production of an utterance, for example *you are* to *you're* (Frank & Jaeger, 2008), the presence of phonetic reduction in predictable contexts and prominence when less predictable (Aylett & Turk, 2004; Bell et al., 2003), in addition to syntactic reduction through the omission of redundant features, such as using the optional English complementizer *that* (Ferreira & Dell, 2000; Jaeger, 2010; Levy & Jaeger, 2007) or case marking (Kurumada & Jaeger, 2015).

However, these studies have focused largely on the way that function words optimize their length for communication, without careful consideration of content words. Even in Piantadosi et al.'s (2011) study, where shorter words were claimed to be used for more efficient communication, no control over the syntactic class of the word was implemented, meaning the results could only suggest generalization across the various word classes. This issue was resolved by Mahowald et al. (2013), who demonstrated through corpus and behavioural studies that content words, that shared similar meanings but varied in length (such as *info* and *information*), would be selected in an utterance based on the speaker's understanding of the context, providing more concise support of Piantadosi et al.'s findings.

If word length is inherently supporting more efficient communication, both for the producer and comprehender, then processes of language change should also reflect this pressure. In order to support this claim, I will return to evidence from the database used by Pagel et al (2007), that assessed rates of replacement in the Swadesh list. Using the database, Monaghan (2014), carried

out further analyses in order to investigate how a variety of psycholinguistic factors can be used to predict rates of replacement. The results showed that, even when controlling for frequency of use, phonological length was a reliable predictor for replacement rate, with longer words being replaced more rapidly than shorter ones. This finding suggests that longer words do indeed undergo dramatic changes over time, with unrelated cognates being introduced into the language at a much faster rate than for shorter words, which like high frequency words, are conserved in the language for longer. This finding suggests, that although there may be increased effort to accommodate longer words comprehension by a speaker, they are still more susceptible to change, with that change being replacement.

Understanding why this might be the case requires an examination of the way that short and long word forms are used within the larger demands of a vocabulary. By assessing the range of possible word forms that can be generated by combining phonetic units of a language together to make a word, if the length of the word is short, for example using three phonemes, then there is a high probability of observing similar sounding words within the set. When compared to words that are longer in length, for example using nine phonemes, this probability decreases dramatically, providing a much larger space for distinctive words to be created. This contrast in the distinctiveness of words has been argued to be an important part of the linguistic system, with the shorter words being used for more predictable meanings (as discussed earlier), allowing for the longer

words to be mapped onto less predictable and more complex meanings (Lewis, Sugarman & Frank, 2014).

If longer words refer to more specific, unpredictable and complex meanings, the demands on the cognitive system to process them accurately will be much greater compared to shorter words, which would be mapped on to more general, predictable and simple meanings. This principle is neatly linked to the classic *word length effect* introduced by Baddeley, Thomson & Buchanan (1975). The effect demonstrated that longer words are recalled less accurately than shorter words. With relation to models of memory, this indicates that longer words are less robustly represented in the brain, whereas shorter words have a much more felicitous representation. Thus, if there is variance in the way short and long words are represented in memory, we may expect to observe variance in the way these words undergo linguistic change, resulting from the same cognitive mechanisms involved in the frequency effect discussed in the previous chapter. In order to investigate whether word length does influence the type of lexical change a word form experiences, the following experiment will again use artificial language learning, with only the length of the word being manipulated. The experiment aims to test whether longer words undergo replacement more frequently than shorter words, replicating Monaghan's (2014) finding. It is also hypothesized, that shorter words will be learnt more accurately, and when errors are made, these errors will be minor adjustments, with the original form still being recognizable.

## Experiment 5: Lexical Stability and Word Length

### Method

**Participants.** The experiment was completed by 21 undergraduate students from Lancaster University (18 female) with a mean age of 18.91 years ( $SD = 0.89$ , range = 18-21). All participants were proficient in English and received course credit for participating.

**Materials.** We used the same set of images for the meaning space as those used in Experiment 4. Likewise, each meaning was paired to a non-word that represented the form of the mapping, generated by using the same consonants and vowels. However, we varied the orthographic length of the forms, using strings that were either 4, 6 or 8 letters long, allowing for a short, medium and long length distinction. In each of the length conditions, there were 4 form-meanings mappings, with all forms adopting a consonant-vowel structure, but varying in the total number of CV syllables. None of the words contained a duplicated CV syllable and any words that resembled English word forms were replaced. Using this same process, 4 different language sets were generated and each participant was randomly assigned a language to learn from during the experiment. An example language can be found in [Table 4.1, p.85](#).

**Procedure.** All participants completed a training phase similar to that used in Experiment 4, the only change being that all form-meaning mappings were presented an equal number of times. This meant that there were still 120 exposures over the course of the training, but each mapping was presented 10

times in total (over a total of 2 blocks). The testing phase was identical to the one implemented in Experiment 4.

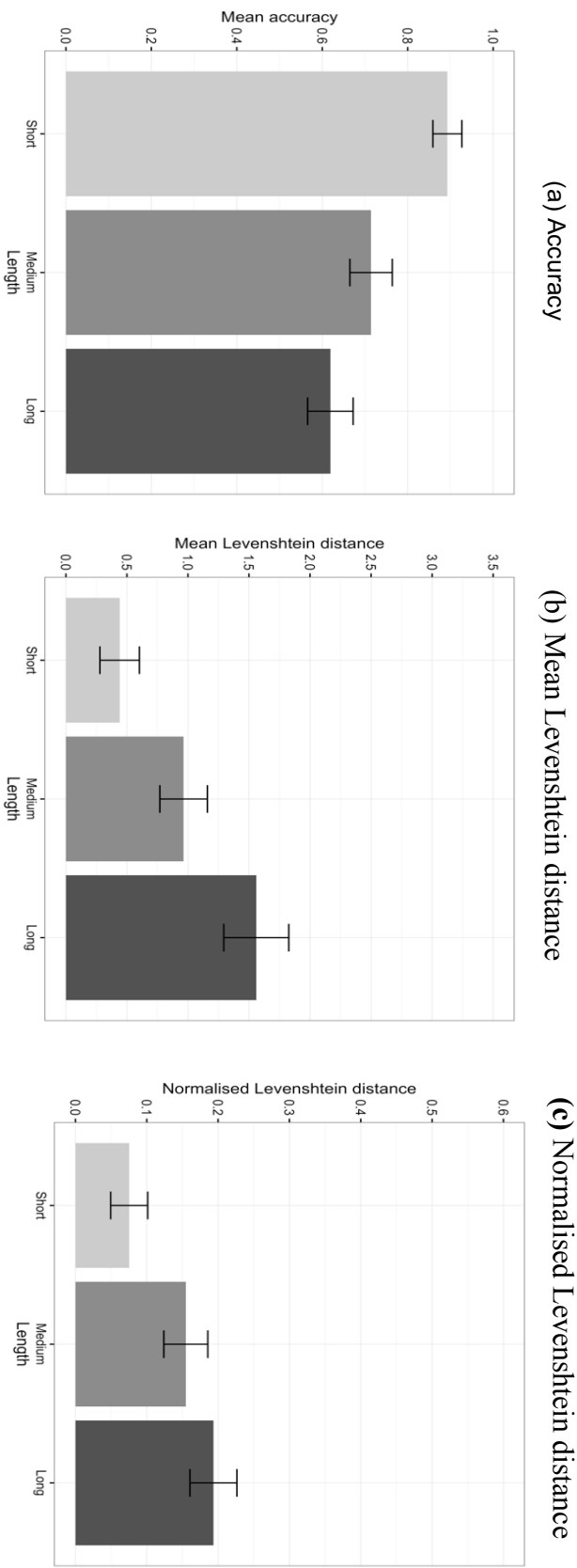
## Results

**Learnability.** We again performed a series of linear mixed-effects models on the results, following the same analysis procedure as we used in Experiment 4, but with length condition as our independent variable, which was again treated as an ordinal factor with levels ordered from short < medium < long.

For the measurement of accuracy, length significantly improved model fit ( $\chi^2(2) = 21.548, p < .001$ ), indicating that accuracy increased as length decreased (estimate = -1.37,  $SE = .32, z = -4.230, p < .001$ ). Additionally, for the Levenshtein distance measure, adding length also showed a significant improvement to model fit ( $\chi^2(2) = 18.156, p < .001$ ), indicating that error decreases as length decreases (estimate = 0.79,  $SE = .18, t = 4.324$ ). This was also the case for the normalised Levenshtein distance measure where the model fit improved by adding length ( $\chi^2(2) = 10.755, p = 0.005$ ), again indicating that as length decreased, error decreased (estimate = 0.08,  $SE = .03, t = 3.24$ ). See [Figure 4.4](#) for results.

### Rates of replacement and adjustment

The data was coded in a way that classified inaccurate responses as either replacements or adjustments. However, as the design of the experiment had word forms of varying length, the critical threshold for the adjustment/replacement distinction was recalculated for each word length. Following the same calculation



**Figure 4.4.** The effects of length condition on learning. (a) shows mean accuracy of participants' recall where lower scores represent lower accuracy, (b) shows mean Levenshtein distance where lower scores represent higher accuracy, (c) shows normalised Levenshtein distance where lower scores represent higher accuracy. Error bars show standard error of the mean by participants (SEM).



procedure as was used in the previous experiment (see [Figure 4.2](#)), we calculated new critical threshold values for the short and long words, on which classifications of adjustments and replacements were made. This was again achieved by computing the average Levenshtein distance between the initial training words and a Monte Carlo sample of 1,000,000 generated words of either 4 or 8 letters (for the short and long words respectively, medium length words retained the same threshold value as was used in the previous experiment as they critical threshold was calculated to be 0.621 (based on  $\mu_{LD} = 6.628$ ,  $\sigma = 1.012$ ,  $len = 8$ ).

Following this, we again performed a series of generalized linear mixed-effects models on our data to see if the fixed effect of length could reliably predict differences in rates of replacement and adjustment, with random effects of subject and training language set. For rates of replacement, length did not significantly improve model fit ( $\chi^2(2) = 4.535$ ,  $p = 0.10$ ). For rates of adjustment, length did significantly improve model fit ( $\chi^2(2) = 96.124$ ,  $p = .001$ ), indicating that the rate of adjustment increased as length increased.

Next, we again wanted to examine if there were significant differences between response types (whether a response was coded as accurate, adjustment or replacement) for word lengths. To achieve this, we followed the same steps as in the previous experiment, constructing a contingency table of counts for each of the response types by length condition, see [Table 4.3](#). Next, using log-linear analyses, a LRT revealed a significant interaction between response type and length ( $\chi^2(4) = 19.922$ ,  $p < .001$ ). However, subsequent analyses revealed that in

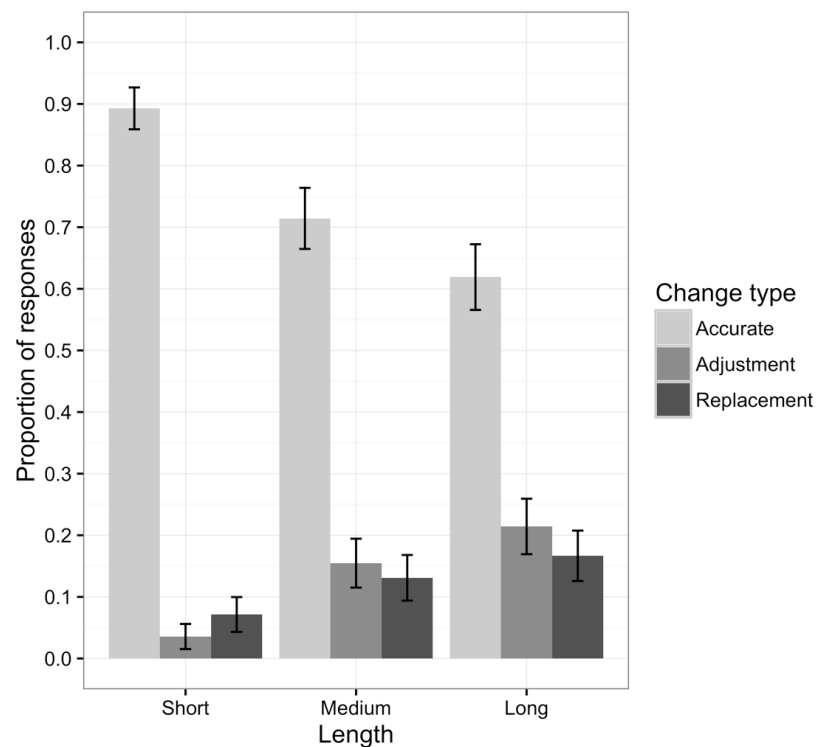
**Table 4.3.** Contingency table showing counts of different response types by length condition.

	Response type		
	Accurate	Adjustment	Replacement
Short	75	3	6
Medium	60	13	11
Long	52	18	14

all word length conditions, there was significantly more accurate responses than both adjustment and replacements (all  $p$ 's  $< .001$ ) and no significant difference between adjustment and replacements in any of the length conditions (all  $p$ 's  $> .05$ ). This indicates that the interaction can be explained by the decrease in accurate responses as length increases, as described previously. See [Figure 4.5](#) for results.

## Discussion

In this experiment we aimed to test predictions that word length affects variation in reliable word production. The results showed that, in line with previous research, the length of a word significantly predicts the accuracy of production for that word. Hence, if a word within a vocabulary exhibits relatively short length, then this will increase its chances of fidelity during use and therefore ensure that the form being used to represent its meaning will be more likely to be conserved in the vocabulary. In contrast, words with a longer length are more vulnerable to production error, which consequently increases their likelihood of



**Figure 4.5.** Proportion of testing responses classified as either accurate, adjustments or replacements in Experiment 5.

change. For example, short words in the language were reproduced relatively well, e.g. *hoku* was recalled accurately, whereas longer words were more likely to have much larger errors, e.g. *gapulona* was recalled as *henumiku* by one participant. This finding would indeed be expected given the previous literature on working memory costs for longer items ([Baddeley et al, 1975](#)).

However, we did find contrasting evidence for our predictions relating to word length and the types of change they are more susceptible to. Whilst we predicted that rates of replacement would increase with word length, we found no statistically significant evidence to support this. Additionally, we found a

significant increase in the rates of adjustment as word length increases, which is counter to our hypotheses. Yet it should be noted that comparisons of rates of replacement and adjustment did not differ within any of the word length conditions, suggesting that these effects may not represent a clear difference between replacement/adjustment rates for different word lengths.

One explanation as to why the rates of adjustment increase with increases to word length could be due to the fact that the present experiment is assessing types of change in an isolated generation of learner, where the changes observed are a single snap-shot of a much more complex process of language change. If there are considerable adjustments made to words with a longer length, then these may over time accumulate to produce more dramatic changes, such as replacements. Given that shorter words are produced more accurately, this would mean that short words would remain stable over this period of time, but longer words would undergo a continuous deviation from their original state, leading to a much more rapid and dramatic process of change, where both adjustments and replacements contribute. Such a view would be consistent with [Monaghan's \(2014\)](#) finding where rates of replacement, over a much longer time scale, are predicted by word length.

Further to this, it is important to note that much of the previous literature on changes in word length, such as reduction (see [Aylett & Turk, 2004](#)) have focused upon spontaneous auditory speech, where the pressures on the producer to make adjustment style changes to the words are proposed to make speech production more efficient. In our present experiment, our word forms are

presented visually as orthographic items. If adjustments in words occur as a response to communicative pressures in spoken language, such effects may be harder to observe in a written medium, where participants may be less inclined to deviate away from the fairly restrictive word structures of consonant-vowel word forms. Thus, it would be essential to explore whether these types of change occur more clearly within a paradigm that requires spoken language, in order to draw more valid and reliable conclusions.

Language change has thus far been argued to have been heavily influenced by two psycholinguistic factors, frequency of use and word length. Whilst these factors are of considerable interest, they do not provide a direct answer to an important question relating to mechanisms of language change: does change occur at the point of acquisition? In order to address this question, it is vital that processes of acquisition are accounted for, if we are to gain a coherent picture of the way a language changes over time.

#### **4.4 Effects of AoA on language change**

Whilst there is undoubtedly strong evidence that demonstrates how language change is a consequence of sophisticated language use (Boyd & Richerson, 1988), there has been increasing interest and demand for a greater understanding of the role that the language learner has in these processes. Christiansen and Chater (2008) put forward a theoretical framework that places the language learner at the centre of language change. In this view, not only the processing constraints of the adult language user are considered as critical factors for

language change and evolution, but also the very fact that languages have to be acquired is considered as a vital pressure, which has important consequences for any theory of language change. If languages are changing and being shaped by the biases introduced from the learner, then this would suggest that language is structured in a way that ensures words that are required at an early stage of language development can be acquired with relative ease.

Given that words are acquired incrementally throughout life, researchers have been interested in developing theoretical explanations and accumulating empirical evidence to investigate the effect of AoA on language processing, with the foundational hypothesis being that there is a processing advantage for early acquired words over words acquired later on in life. To demonstrate this, we can return to our examples of the word *'the'*, which is acquired relatively early on in life and is processed with high accuracy and speed. However, this contrasts with the word *'mizzenmast'*, which is acquired much later and does not exhibit the same processing advantages in comparison. Whilst there is a difference between the AoA of these words, there are also inevitable inter-correlations with other psycholinguistic properties of the words, most notably frequency (Zevin & Seidenberg, 2004 reported a correlation of  $-.71$  between AoA and subjective frequency ratings). This understandably may confound any possible effects of AoA. Therefore, any well attested account that places importance on AoA, must also demonstrate that this effect is independent from other potentially confounding factors.

Indeed, there are convincing theoretical arguments for why an AoA effect should and does exist in its own right. There are currently two main models that are used to explain how such an AoA effect may operate (although see [Johnson & Barry, 2006](#); [Juhasz, 2005](#) for a wider review of other possible accounts). Firstly, [Ellis and Lambon Ralph \(2000\)](#) established the influential framework that the AoA effect is a result of changes to the plasticity of the system involved when learning new mappings. Using a connectionist model, they demonstrated that when learning occurs incrementally, the words that are acquired early on during training can optimize their own connection weights. This is argued to be a consequence of the system's increased plasticity at early stages of acquisition, when learning is not constrained by an established set of connection weights. Once words acquired early on have established such a weighting preference within the system, then any additional words that are introduced into this system must be accommodated in a less plastic network, therefore have less control over the weightings.

An alternative suggestion for the AoA effect specifies that it is semantics that is the locus of the explanation. The importance of semantics in the AoA effect was initially proposed by [Van Loon-Vervoornt \(1989\)](#) and then notably developed by [Brysbaert, Wijnendaele and De Deyne \(2000\)](#). This view holds that the incremental organization of the semantic system ensures that later acquired concepts are built on to earlier acquired ones, this results in later acquired words taking a more indirect processing path as they have a less central role in the semantic system, unlike those words acquired earlier in life where

access is more direct, given their semantic prominence. Evidence for this has come from [Steyvers and Tennenbaum \(2005\)](#), who developed a semantic growth model, where early acquired words are shown to be the central connection to other later acquired words, that have comparably fewer semantic connections. In subsequent experimental studies (see [Brysbaert & Ghyselinck, 2006](#); [Catling & Johnston, 2009](#)), the role of semantics in the AoA effect has been demonstrated to be facilitative in the reported magnitude of the effect.

Indeed, there is strong evidence from empirical investigations into how AoA, which is often assessed using measurements from subjective ratings ([Kuperman, Stadthagen-Gonzalez & Brysbaert, 2012](#) for example) highlight a processing advantage for early acquired words over words acquired later in life across a broad variety of domains, such as lexical decision, object naming and reading (see [Johnston & Barry, 2006](#); [Juhasz, 2005](#) for review).

Whilst evidence for the AoA effect appears robust, there has recently been more focused attention towards exploring the effect of AoA on word retention and loss. For instance, [Navarrete, Pastore, Valentini and Peressotti \(2015\)](#) demonstrated that experiencing a tip-of-the-tongue state (where a word is temporarily inaccessible during lexical retrieval) is more likely when the word is acquired later in life. Additionally, [Marful, Gómez-Ariza, Barbón and Bajo \(2016\)](#) showed that early acquired words were more resilient to instances of forgetting, whereas late acquired words appeared more susceptible. Relatedly, [Belke, Brysbaert, Meyer and Ghyselinck \(2005\)](#) put forward the argument that early acquired words act as strong lexical competitors to their semantically



related, but later acquired, counterparts. This would mean later acquired words are weakened during lexical selection and subsequent production.

Likewise, within various populations of patients with neuropsychological disorders or cognitive impairments, there have been substantial links made between AoA and performance on various psycholinguistic tasks. This body of research highlights the fact that early acquired words appear to have a privileged position in the vocabulary that allows them to endure such deficits; unlike later acquired words that are seemingly more vulnerable to loss (see [Brysbaert & Ellis, 2015](#); [Ellis, 2012](#) for review). These findings are also consistent with results from healthy elderly participants, where early acquired words have also been shown to be retained much longer in life than late acquired words ([Hodgson & Ellis, 1998](#)), again providing further evidence for the proposition that early acquired words are more resilient to loss within the cognitive system.

Thus, there is substantial evidence in the AoA literature that demonstrates the processing advantages of early acquired words over those acquired later in life, in addition to work that suggests AoA can also be linked to whether a word is retained, forgotten or chosen during production. However, when connecting these findings to processes of language change, it would appear that any causal link made between the two would be ultimately indirect. In order to overcome such an issue, [Monaghan \(2014\)](#) has provided evidence that demonstrates a quantifiable difference between the rate at which early and late acquired words are replaced in a language. Using the same data as [Pagel et al. \(2007\)](#), [Monaghan](#) showed that late acquired words undergo more rapid rates of replacement than

early acquired words, even when controlling for other psycholinguistic variables such as frequency, word length and concreteness.

Although the finding from [Monaghan \(2014\)](#) provides an interesting insight into how AoA and language change may be intricately connected, the results could benefit from additional support. It could be argued that the use of the Swadesh list, a list of only 200 fundamental items, may only provide a narrow snapshot of natural language's full lexical inventory. Moreover, AoA norms are collected from modern day ratings, and there can be no certainty that these ratings would have adopted the same pattern of acquisition over the course of thousands of years. This may mean that words acquired by infants learning a language today may not be identical to the semantic meanings being learnt by infants many years ago, given the flexibility in the evolution of semantics ([Urban, 2011](#); [Winter, Thompson & Urban, 2013](#)).

In order to address these issues, whilst also isolating AoA effects more generally (see [Lewis, 2006](#)), an alternative approach to studying such AoA effects has been offered by laboratory based analogues. With the supposition being that if AoA effects are to be considered a fundamental part of the way humans learn (be it languages or other behaviours), then such effects should be observable under controlled laboratory conditions. Importantly, such a claim has received a growing amount of support, with a number of studies successfully demonstrating AoA effects using experimental paradigms that afford the experimenter control of other potentially confounding factors. For instance, [Stewart and Ellis's \(2008\)](#) study showed an AoA effect for the categorisation of

checkerboard patterns by manipulating which patterns were shown early and late during stimulus exposure. Furthermore, such a paradigm has been used to show effects using linguistic stimuli, with effects reported from designs using words from natural language (Izura et al, 2011; Tamminen & Gaskell, 2008) as well as from artificial languages where novel word forms were presented and learnt (Catling, Dent, Preece & Johnston, 2013; Joseph, Wonnacott, Forbes & Nation, 2014).

To test further the finding from Monaghan (2014), that early acquired words undergo much less lexical replacement than late acquired words, the aim of the next experiment was again to investigate such a finding through the use of a laboratory analogue. By manipulating the order in which words within an artificial language were presented during training, the aim was to observe if there is any difference in accuracy between how early and late acquired words are reproduced. Further to this, using the same calculations as used in the previous two experiments, we aimed to investigate whether rates of replacement also differ. One additional hypothesis to be tested will relate to rates of adjustment in the participant's production output. Whilst there is no prior evidence that demonstrates a quantitative difference for AoA and adjustments in language, or indeed any established theoretical framework that explores such an issue, we would predict that, as was predicted for the frequency and length hypotheses, early acquired words would be more likely to undergo adjustment and not replacement when errors are produced.

## Experiment 6: Lexical Stability and AoA

### Method

**Participants.** The experiment was completed by 25 undergraduate students from Lancaster University (22 female) with a mean age of 19.44 years ( $SD = 3.29$ , range = 18-34). All participants were proficient in English and received course credit for participating.

**Materials.** We used the exact same sets of form-meaning mappings as those used in Experiment 4. To investigate the effect of AoA on participants' learning, we assigned each of the mappings to either an early or late acquired condition, with each condition containing 6 form-meaning mappings. Mappings would be presented with varying weightings during training, depending on which AoA condition they were in.

**Procedure.** All participants completed a training phase similar to that used in Experiments 4 and 5. However, exposure to the mappings were weighted to ensure that non-words that were in the early acquired group would be presented predominately early on during training, whilst late acquired words would be presented predominately later on. This was achieved by varying the number of exposures either early or late acquired mappings would have in each of the training blocks (See [Table 4.4](#) for exposure weightings in each block), such a design has been used in previous laboratory based analogues investigating effects of AoA, such as [Stewart and Ellis \(2008\)](#). In order to ensure that the early acquired words were acquired early on in training, participants completed a short pre-test block immediately after the first training block, this pre-test was identical

**Table 4.4.** Training and testing procedure for Experiment 6, showing number of exposures each form-meaning mapping was given at each block of training.

Block	Acquisition	
	Early	Late
1	6	0
----- Pretest phase		
2	1	3
3	1	3
4	1	3
5	1	1
----- Testing phase		
Total	10	10

to the testing phases used in the previous experiments, but only the early acquired words were tested.

By using such a design, we ensured that all mappings would be given an equal 10 exposures each during the training phase, giving a total of 120 training exposures over the 5 blocks. In order to reduce any recency effects (following [Izura et al, 2011](#)) the final block of training had equal weightings for the early and late acquired mappings. There was a final testing phase once all block of training had been completed. This was identical to the one used in the previous experiments in this chapter, where all form-meaning mappings were presented randomly. An example language can be found in [Table 4.1, p.85](#).

## Results

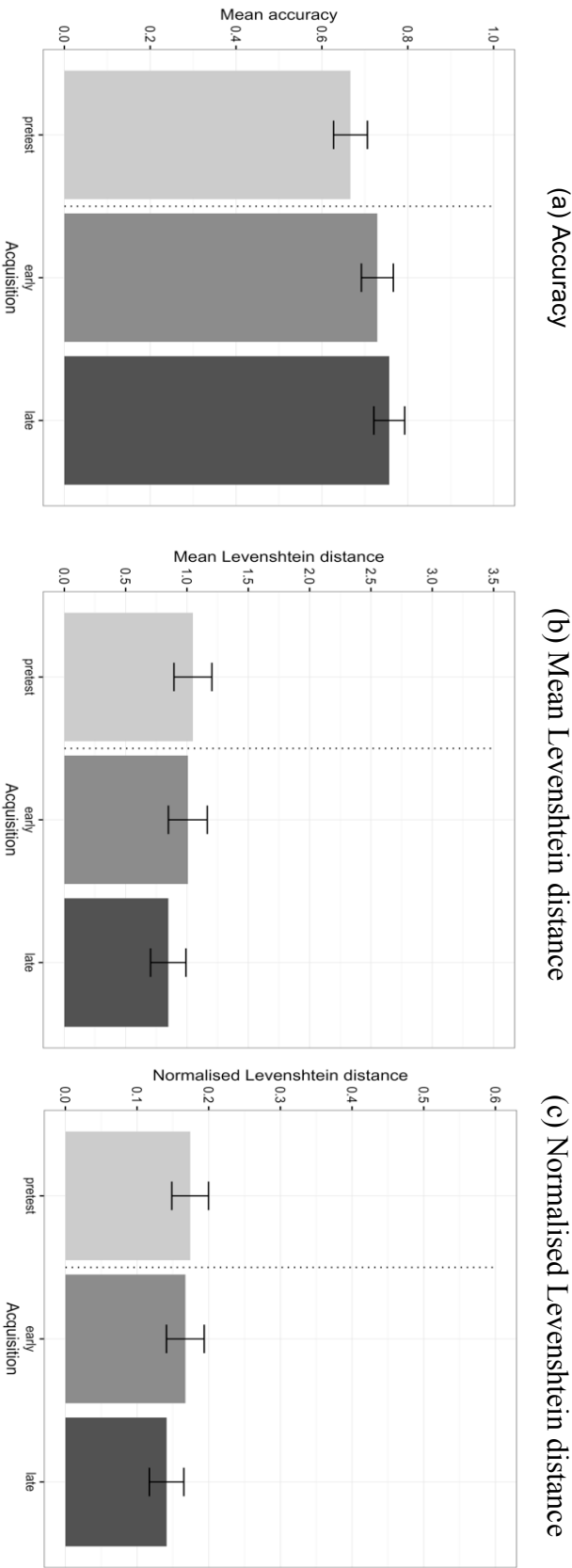
To ensure that participants were acquiring the mappings of the early acquired set early on in training, we analysed the performance from the pre-test phase for

accuracy. Performance on this pre-test phase demonstrated that participants were on average reproducing the trained words accurately, with the mean normalised Levenshtein distance of 0.17 (SD = 0.31). Whilst there was no available comparison to chance, this result suggests that on average participants were recalling words with more than 80% accuracy. Given that performance in Experiment 4 for the high frequency condition showed participants produced an average normalised Levenshtein distance of 0.16 (SD = 0.31), we interpret the pre-test results to represent sufficient learning.

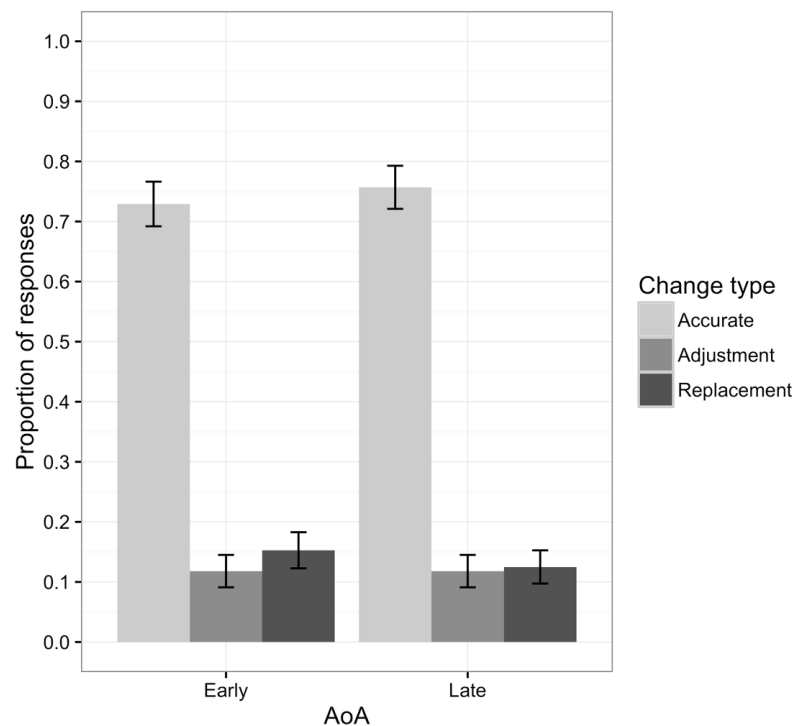
**Learnability.** We performed a series of generalized linear mixed-effects models on the data, following the same analysis procedure as we used in Experiments 4 and 5, but with AoA as our independent variable. AoA was treated as a categorical factor with comparisons being made between early and late acquired groups.

For the measurement of accuracy, AoA did not significantly improve model fit ( $\chi^2(1) = 0.382, p = 0.54$ ). Likewise, the Levenshtein and normalised Levenshtein distance measures also showed no significant improvement to model fit when adding AoA ( $\chi^2(1) = 0.849, p = 0.357$  and  $\chi^2(1) = 0.818, p = 0.366$ ). See [Figure 4.6](#) for results. Thus, the results show no significant difference between the early and late acquired groups for performance during production.

**Rates of replacement and adjustment.** As the word forms used in the present experiment were the same as those used in Experiment 4, the same critical threshold was used to determine adjustments/replacement, see [p.102](#) for



**Figure 4.6.** The effects of AoA condition on learning. **(a)** shows mean accuracy of participants' recall where lower scores represent lower accuracy, **(b)** shows mean Levenshtein distance where lower scores represent higher accuracy, **(c)** shows normalised Levenshtein distance where lower scores represent higher accuracy. Error bars show standard error of the mean by participants (SEM).



**Figure 4.7.** Proportion of testing responses classified as either accurate, adjustments or replacements in Experiment 6.

details. We performed a series of generalized linear mixed-effects models on our data to see if the fixed effect of AoA could reliably predict differences in rates of replacement and adjustment, with random effects of subject and training language set. For rates of replacement, including the AoA term did not significantly improve model fit ( $\chi^2(1) = 0.680, p = .41$ ), indicating that there was no difference between the rates of replacement and the AoA of the word being presented during training. Similarly, for rates of adjustment, AoA did not



significantly improve model fit ( $\chi^2(1) = 0, p = 1$ ), indicating that there was again no difference between the rates of adjustment and AoA. See [Figure 4.7](#) for results.

## Discussion

In the present experiment we aimed to test the prediction that early acquired words would experience fewer production errors and lower rates of replacement than words acquired later in the training. The results revealed no significant differences between the early and late acquired words for any of our dependent variables. Whilst these findings do not provide evidence in support of our original hypotheses, or indeed results that are in line with previous research investigating the AoA effect, there may be alternative explanations for the results reported here.

Firstly, it should be noted that there were considerable differences in the design of the present study and those that have previously reported AoA effects in the laboratory. Primarily, this difference is in the training regime implemented, which is how AoA is manipulated. Although in the present experimental design we weighted the number of exposures for early and late acquired words over the course of training, which is consistent with previously used designs (such as in [Catling et al, 2013](#); [Izura et al, 2011](#); [Stewart & Ellis, 2008](#)), the training given to participants in our study did differ in terms of the way the training blocks were

scheduled. Previous experiments were run over the course of several days, with participants completing each training block on separate days, which contrasts to our design where all blocks are completed during the single experimental session, i.e. on the same day. By not using a design where sufficient incremental consolidation of the early acquired words occurs, then a core theoretical aspect of the AoA effect (see [Brysbart et al, 2000](#); [Ellis & Lambon Ralph, 2000](#)) is not being introduced within the experimental design. Therefore, it may be that both early and late acquired words are being acquired in a similar way in the current study, without any sufficient distinction made by the participant.

Secondly, the present experiment used word forms that were arbitrarily paired to a visual meaning, whereby the form had no clear relationship to the meaning. However, recent empirical studies have shown that words that are acquired early on in life are in fact incorporating sound-symbolic properties within their forms, thus early acquired words may not necessarily be arbitrarily related to their meaning. [Monaghan et al. \(2014\)](#) demonstrated that early acquired words were in fact significantly more systematic than words acquired later in life, whilst [Perry, Perlman and Luypan \(2015\)](#) reported high rates of iconicity in early acquired words. Although previous reports of the AoA effect have not considered the importance of sound-symbolism to the effect, yet still found marked differences between how early and late acquired words are processed, it appears that the role of sound-symbolism for the acquisition of words early on in life is steadily gaining greater theoretical importance. Therefore, it may be plausible

that without incorporating such sound-symbolism within the forms of the early acquired words, the hypothesised AoA effect did not materialise.

## 4.5 General Discussion

Three experiments were carried out with the aim of generating previously reported effects that demonstrate how psycholinguistic properties of a language can be used to reliably predict differences in the way words may undergo change (Monaghan, 2014; Pagel et al, 2007). Whilst these previous findings have relied upon methods and data which, although providing an intriguing insight into how words change, may not provide a completely reliable account on which to draw robust conclusions. Instead, we have here a novel experimental methodology and analysis to provide more reliable evidence, that will further our understanding of the way psycholinguistic properties play a critical role in the way language changes.

We find that both word frequency and length can be used to reliably predict where production errors will occur, results that are consistent with previously reported frequency and length effects, although we did not find any significant effect for AoA. However, by using an artificial language learning paradigm to test our hypotheses, and therefore ensure strict control over our variable of interest, we can be sure that these effects are independent of one another. This is of particular importance given the strong inter-correlation between word frequency and length, as shown by Zipf (1949).

For our analyses regarding rates of replacement, we only found a clear difference for the frequency experiment, not for length or AoA, whilst for rates of adjustment, we found no evidence in support of our original hypotheses. It would be important to consider the role of communicational context as an explanation for these results. Much of the research that demonstrates adjustments occur when comprehension is understood to be fairly reliable, which allows the speaker to incorporate minor changes to word forms, i.e. if a word has a high likelihood of being processed accurately, then a pressure for more efficient production may be introduced ([Zipf, 1949](#)). If our experimental paradigm is to produce a valid analogue for which claims about language change are to be drawn, then the theoretical importance of having a production test that is situated within a communicative context, where the demands of both producer and receiver are considered, must be incorporated.

Likewise, it is also important to consider the way natural language is learnt. An adult can acquire a very large and varied vocabulary, which contains words that are highly frequent, short and acquired early, yet we are still able to acquire words that are very infrequent, long and acquired later in life. Whilst the experiments presented in this chapter capture some aspects of learning, a possible confound is that the words that have high rates of error (most notably the low frequency words), may not have been learnt at all. This may reduce the extent to which the results can generalise to how natural language is learnt, and consequently, changes as a result of cognitive constraints.

Furthermore, the present set of experiments offer insights into the way a single participant's biases for either frequency, word length or AoA may contribute to processes of language change. In order to provide a reliable account of such processes, one must acknowledge that languages change, not simply as a result of a single person's learning, but instead over time through the cumulative cultural evolution of the language (Kirby, 2000). Given that small biases introduced by a learner can be magnified through the transmission of information over several generations of learners (Griffiths, Kalish & Lewandowski, 2008), we may expect to see how small changes introduced by a learner (as shown in the present set of experiments) can gradually change, or indeed conserve, properties of a language over time. This will be the focus of the next chapter.

# Chapter 5

## Predictors of lexical stability through the cultural transmission of language

---

### 5.1 Introduction

Traditionally, artificial language learning experiments have been used by cognitive scientists to explore how certain psycholinguistic properties of language can influence aspects of processing and production in a controlled laboratory setting. However, when one is considering questions relating to how psycholinguistic properties can play a role in language change and evolution, then such paradigms can only provide a snapshot of how an individual learner may contribute to changes. Yet this is far from what actually happens during the process of language evolution. Instead, languages (like many other cultural traits) are transmitted across generations of learners, where they are continuously being learnt, used and then passed onto the next generation of learners, with small changes cumulatively being introduced, leading to the emergence of more

significant changes (Kirby, 2001; Griffiths & Kalish 2007). This process of iterated learning has been widely accepted as a critical model on which to base accounts of language evolution (Christiansen & Chater, 2008).

Researchers have been utilising the iterated learning model to explain the emergence of several key properties of human language through the use of computational and mathematical modelling, and also behavioural experiments with human participants (see Kirby, Griffiths & Smith, 2014 for review). A core finding from these studies is that as a result of the language being culturally transmitted, there will be an increase in the learnability of the language, such that error between generations of learners decreases over time (Kirby, Cornish & Smith, 2008). This finding has been attributed to the manifestation of learning biases shared by the language learners, which shape the evolution of the language once it has been acquired. Thus, the cultural evolution of language through iterated learning has been used more recently to explain earlier findings, which have demonstrated a surprising number of shared universals across languages (Comrie, 1981; Greenberg, 1963).

Whilst there has been much work using the iterated learning model to explore which of the universal properties of language emerge as a result of cultural evolution, there has been relatively little work that uses it to investigate which properties of language are conserved, or indeed the differences in the rate at which the properties undergo change. If the iterated learning model is to account for the emergence of shared universals, then it should also account for the rich diversity in language too (Croft, 2000; Evans & Levinson, 2009). This

point is highlighted by [Rafferty, Giffiths and Ettlinger \(2013\)](#), who provide counter-examples to the claim that greater learnability is sufficient for a property to become prevalent in a language. For instance, within a list of concepts, a certain item may be more distinct from the others (e.g. the word *elephant* within a list of grocery items), whilst this item may be easier to learn, the process of iterated learning does not lead to it dominating the list, indeed once it has been lost from the list, then it would be unlikely to spontaneously reappear.

As discussed in [Chapter 4](#) (see also [Monaghan, 2014](#)), there is growing evidence that links the way words are processed (based on their psycholinguistic properties) to the rates at which the word will undergo change. However, the manipulation of psycholinguistic properties found in natural language, within many iterated learning studies and their influence on the way the language evolves, has thus far yet to be comprehensibly investigated. There is evidence to suggest that when psycholinguistic factors have been explored within iterated learning models, then interesting results can emerge that do reflect natural language more closely. For instance, in a computational model of iterated learning by [Kirby, Dowman and Griffiths \(2007\)](#) (see also [Kirby, 2001](#)), frequency was varied during training to reflect the Zipfian ([Zipf, 1936](#)) distribution found in natural language. What emerged from this model was a dynamic language that underwent change constantly over the course of transmission, reflecting the non-stationary nature of natural language. Interestingly, the manipulation of frequency also led to the emergence of regularity in low frequency words, but irregularity in high frequency words,



consistent with other accounts of regularization in language (Francis & Kučera, 1982; Lieberman et al., 2007). However, the manipulation of frequency within behavioural analogues has yet to be reported, with Kirby et al (2008) using a uniform frequency distribution in their experimental paradigm. In other behavioural studies of iterated language learning, Lewis and Frank (2015) demonstrated that word lengths would adapt to suit the complexity of the meaning being represented, revealing that word length can play an important role in the way that languages evolve. Critically though, such studies have aimed to investigate whether these properties of language emerge as a result of the cultural transmission process, but not whether these properties are retained or lost from the outset.

Observing changes in languages over the course of hundreds or thousands of years has also posed difficulties for researchers. The documentation of languages is understandably limited, but phylogenetic studies of language change have provided promising insights into how languages may have diverged over time (Gray & Atkinson, 2003). Building on this work, the field of cladistics has aimed to explain difference across different languages, but also when similarities may also be found (Atkinson et al., 2008; Blasi et al., 2016; Pagel et al., 2007). Adopting such an approach offers researchers the opportunity to reconstruct the history of a language and calculate estimations of when certain elements within human languages have undergone change. This in turn, allows for a quantitative investigation into which factors may be driving such changes, or indeed explaining when little change occurs (Pagel et al., 2007; Pagel & Meade, 2006).

In order to assess the way that lexical items in the vocabulary undergo differing rates of change over a long timescale, [Pagel and Meade \(2006\)](#) calculated estimates of how often cognates have been replaced by an unrelated form for words from the Swadesh list of fundamental vocabulary items ([Swadesh, 1952](#)). These estimates for how often a word will be replaced have since been used in regression analyses with psycholinguistic properties being used to predict differences in these rates of replacement, with significant results reported for frequency ([Pagel et al, 2007](#)), word length and AoA ([Monaghan, 2014](#)) and semantic factors such as number of synonyms, number of senses and imageability ([Vejdemo & Hörberg, 2016](#)). Although these findings offer intriguing results, their validity remains contentious. Considering that the Swadesh list is comprised of only 200 lexical items, the scope of the vocabulary covered by these analyses is significantly restricted, this also limits the range of the psycholinguistic represented, with only fundamental vocabulary items being represented.

The aim of this chapter is therefore to explore using an experimental methodology the results reported from previous cladistic studies, where evidence suggests that psycholinguistic properties play an important role in the way that a language changes over time. But here, we present a novel way of testing this prediction using a well-established model of the cultural transmission of language, with the benefit of controlling stimuli and variables under laboratory conditions. We hypothesise that the properties of frequency, word length and AoA will act as reliable predictors for the rates of error, adjustment and

replacement in the language, but with the added dynamic of transmission across generations of learners, which will develop on the findings from [Chapter 4](#).

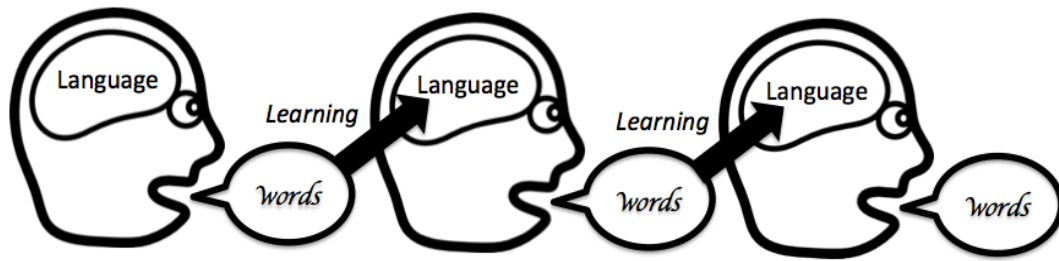
## 5.2 Experiment 7: Frequency and iterated learning

### Method

**Participants.** The experiment was completed by 32 students from Lancaster University (20 females) with a mean age of 21.4 years ( $SD = 2.70$ , range = 18-30). All participants were proficient in English and received £3.50 for participating.

**Materials.** The initial starting languages presented during training to the first generation in each transmission chain were identical to those used in [Chapter 4's frequency](#) experiment. This meant that the same consonants and vowels were used for the word forms, as well as the same 12 visual stimuli to represent the meanings (see [p.83](#) for details on their design and [Table 4.1, p.85](#) for examples of the language). Words were generated through a pseudo-random process, ensuring that at the start of the transmission chain, there was no underlying correspondence between the phonology of a word and its combinatorial structure.

**Procedure.** The experiment involved learning and then reproducing an artificial language, whereby the frequency of the words were manipulated (see [Chapter 4, p.84](#) for details on how these were manipulated). However, participants in the experiment would learn from an initial training language



**Figure 5.1.** Visual representation of the iterated learning model, taken from [Roberts \(2013\)](#). Each person represents a generation, each of which learns a language based on the linguistic output of the previous generation.

created by the experimenter, but then generate a new language that the subsequent participant would learn from. In order to achieve this, an iterated learning paradigm was used. This involved chains of learners who would transmit the language across generations.

Within such a design, the first participant in the chain was presented one of the initial starting languages during a training phase, which they were then asked to reproduce during a testing phase immediately after the training phase had been completed. The responses that the participant produced in the testing phase generated a new set of form-meaning mappings. This new set was then used as the language to be presented to the next participant in the chain. This process was then repeated until the end of the transmission chain had been completed (see [Figure 5.1](#) for an example of how this process works). In each experiment there were 4 independent chains, each with 8 participants, each of the chains began with a different initial starting language. None of the participants

were informed about this process until they were debriefed after the experiment had been completed. It should also be noted that chains were run in parallel to each other, meaning that chains were not completed consecutively or all on the same day.

Previous experiments using iterated learning to investigate language evolution have demonstrated that underspecification may emerge within the language, where the same word form will be used to represent several meanings (see Experiment 1 of Kirby et al., 2008). As underspecification results in a language that does not meet the communicative needs of natural language, i.e. to be communicatively expressive and not dominated by ambiguity. Therefore, an artificial filtering process was carried out before the next generation's training commenced to eliminate such underspecification. This was similar to the process implemented in Experiment 2 of Kirby et al. (2008), whereby the experimenter would assess if any word form was produced more than once during the testing phase of the experiment. When this did occur, one of the words was chosen at random and kept in the language, any other words that were duplicates were assigned a new word form. This new word form was chosen randomly from the inventory of words from the other initial starting languages. This process was also used when there was no response produced by the participant, which happened only once.

**Analysis.** The analyses aimed to investigate how the language generated during the testing phase of the experiment differed from the language the participants were exposed to during the training phase. This involved the analysis

of four main dependent variables: accuracy, error distance, number of adjustments and number of replacements (this was the same as in [Chapter 4](#), with the variables being calculated in the same way).

Accuracy was again quantified by assessing whether a response was reproduced identically to the word presented during training, if the response was identical it was coded as accurate (or ‘1’ in the dataset), else inaccurate (or ‘0’ in the dataset). For the error distance, the normalized Levenshtein distance was calculated, which was the minimum number of insertions, deletions and substitutions required to transform the word presented during training into the word produced during testing. This value was then normalized by dividing it by the length of the longest word, ensuring that a standard measurement can be used for all experiments and all word lengths. Responses that had less error received a distance measurement tending towards 0, whilst inaccurate responses tended towards 1.

Adjustments and replacements were again classified based on the error measurements and whether they were above (for a replacement) or below (for an adjustment) the critical threshold value (as described on [p.88](#)). This remained 0.590 for all words that were 6 characters long (which was the case for all words in the present experiment). This allowed us to quantify the number of adjustments and replacements occurring at each generation within the chain of learners, therefore we could observe how these rates might change over the course of transmission. This was coded as two separate variables, with

adjustments/replacements being represented as ‘1’, with ‘0’ being used to represent all other remaining responses.

We analysed these variables using mixed-effects models, with the fixed effect of either frequency, word length or AoA, depending on the experimental manipulation. We also included generation in to the models, from 1 (the first participant) to 8 (the last participant). The following analyses report results from Likelihood Ratio Tests (LRT) (Baayen, 2008). This involved stepping through a series of models that have been built up from an initial model containing just random effects of learning chain and word meaning, which is then proceeded by subsequent models that incrementally introduce the fixed effects. Model comparisons were made to assess whether the inclusion of the fixed effect significantly improved the fit of the model, with models being compared to the best fitting model that did not include the additional fixed effect.

Below, we report the results from the three different experiments individually, addressing each of the dependent variables in turn.

## Results

### Learnability

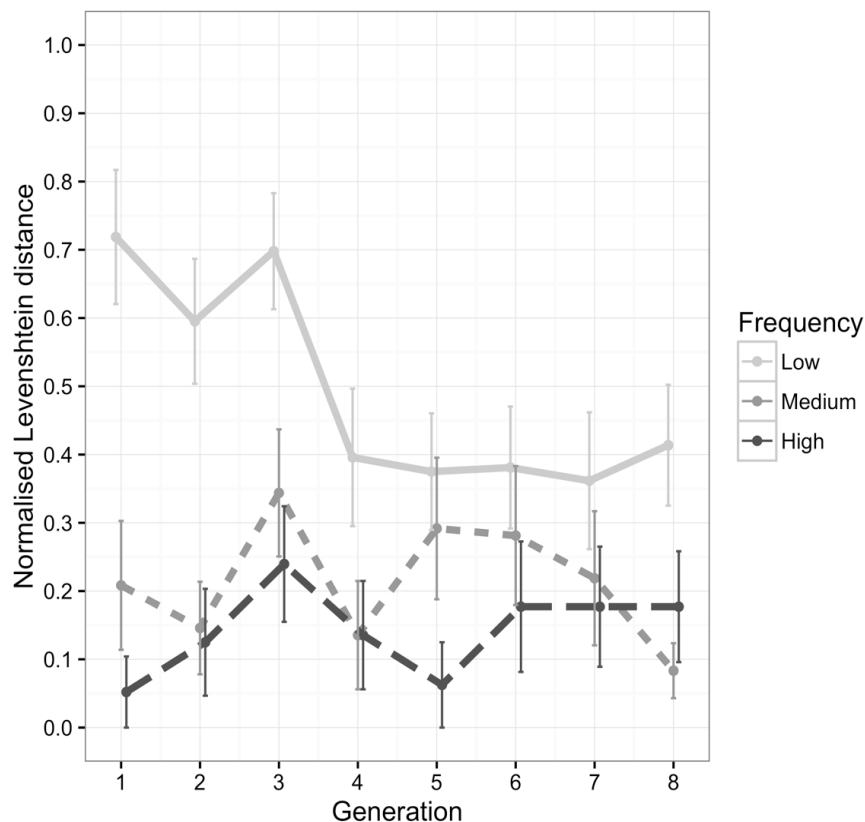
*Error.* For the dependent variable of normalised Levenshtein distance (a measurement of string similarity between input and output), results from the LRTs revealed that on the full data set, the addition of frequency significantly improved model fit ( $\chi^2(2) = 69.795, p < .001$ ), indicating that there was a significant difference in the amount of error participants made during testing,

whereby increases in frequency resulted in less error. Further to this, the addition of generation to the model also improved the fit ( $\chi^2(1) = 4.363, p = .037$ ), indicating that over the course of transmission error significantly decreased. Importantly, the addition of the interaction term frequency x generation also significantly improved model fit ( $\chi^2(3) = 12.196, p = .002$ ), indicating that as the language was transmitted across generations, there was a significant change in error as a result of frequency being varied.

In order to explore this interaction, we then conducted follow up analyses on the subsetting data from each of the frequency conditions. For the low frequency trials, we found a significant improvement to model fit when generation was added to the model ( $\chi^2(1) = 13.828, p < .001$ ), demonstrating that error was decreasing over the course of transmission (estimate = -.05,  $SE = .01$ ,  $t = -3.836$ ). This was not the case for the medium or high frequency conditions, where no significant improvement to model fit was found ( $\chi^2(1) = 0.237, p = .626$  and  $\chi^2(1) = 0.845, p = .358$  respectively), indicating that there was no significant changes in the amount of error produced when frequency is increased. See [Figure 5.2](#) for results and [Table 5.1](#) for an example of how the language changed from generation1 to 7.

*Accuracy.* Results from the LRTs revealed that on the full data set, when assessing the number of accurate responses the addition of frequency significantly improved model fit ( $\chi^2(2) = 75.292, p < .001$ ), indicating that as frequency increased the number of accurate responses also increased (estimate = 1.67,  $SE = .22$ ,  $z = 7.63$ ). However, the addition of generation to the model did





**Figure 5.2.** Mean error of production by generation in the transmission chain for Experiment 7. Values were calculated from the normalised Levenshtein distance between the training and production words for each of the frequency conditions. Error bars show SEM.

not improve the fit ( $\chi^2(1) = 0.916, p = .339$ ), nor did the addition of the interaction term frequency x generation ( $\chi^2(3) = 5.130, p = .163$ ), indicating that there was no significant changes in accuracy as the language was transmitted across generations, or that there was any change in accuracy across generations as a result of the frequency of the words varying.

In follow up analyses however, there was a marginally significant improvement to model fit in the low frequency condition when generation was added to the model ( $\chi^2(1) = 3.476, p = .062$ ), suggesting that accuracy was

**Table 5.1.** An example of the word forms produced at each generation of Chain C, by frequency condition.

Frequency	Initial	G1	G2	G3	G4	G5	G6	G7	G8
Low	lemopa	wopehu	migowe	nupowe	nupowe	nupowe	nupomu	muphonu	gihomu
Low	galeki	hupila	wuhowe	mugowe	mupuwe	mipuwe	mupila	gianho	gianho
Low	nokali	pinamo	haguno	hupuwe	hupuwe	mipuwe	gihomu	nupiamo	giano
Low	mihuge	nuhopi	migowe	nupowe	ligepe	ligepe	muhopi	gihano	gianhu
Medium	hewino	nugohu	nugohu	nugohu	nugohu	nugohu	nugohu	gihamo	gihamo
Medium	kigowu	nugowu	nugowu	nupuwe	nupuwe	nupuwe	muhopi	wamoke	wamoke
Medium	gonime	nugohu	mipuwe	mipuwe	mipuwe	mipuwe	muhopi	gihamo	gahemi
Medium	wopehu	wopehu	hakupo	hakupo	hakupo	hakupo	ginohu	ginohu	gianhu
High	nakuwo	nuponu	nuponu	nuponu	nuponu	nuponu	nuponu	nuponu	nipewa
High	hupila	hupila	hupila	hupila	hupila	hupila	hupila	hupila	hupila
High	pinamo	pinamo	pinamo	pinamo	pinamo	pinamo	pinamo	pinamo	pinamo
High	muhopi	muhopi	muhopi	muhopi	muhopi	muhopi	muhopi	nuponu	nuponu

improving marginally over the course of transmission (estimate = .16,  $SE = .09$ ,  $z = 1.834$ ). This was not the case for the medium or high frequency conditions, where no significant improvement to model fit was found ( $\chi^2(1) = 0.453$ ,  $p = .501$  and  $\chi^2(1) = 1.262$ ,  $p = .261$  respectively), indicating that the number of accurate responses remained stable even when the language was being transmitted across generations. See [Figure 5.3](#) for results. Whilst this analysis of accuracy provides insights into how many responses were produced with complete fidelity (i.e. a faithful recall based on the training input), the Levenshtein distance (used in our error analysis) offers a considerably more sensitive measurement of change, where small and large changes are accounted for. This also provides the basis for the next part of our analyses, where rates of replacement and adjustment will be presented.

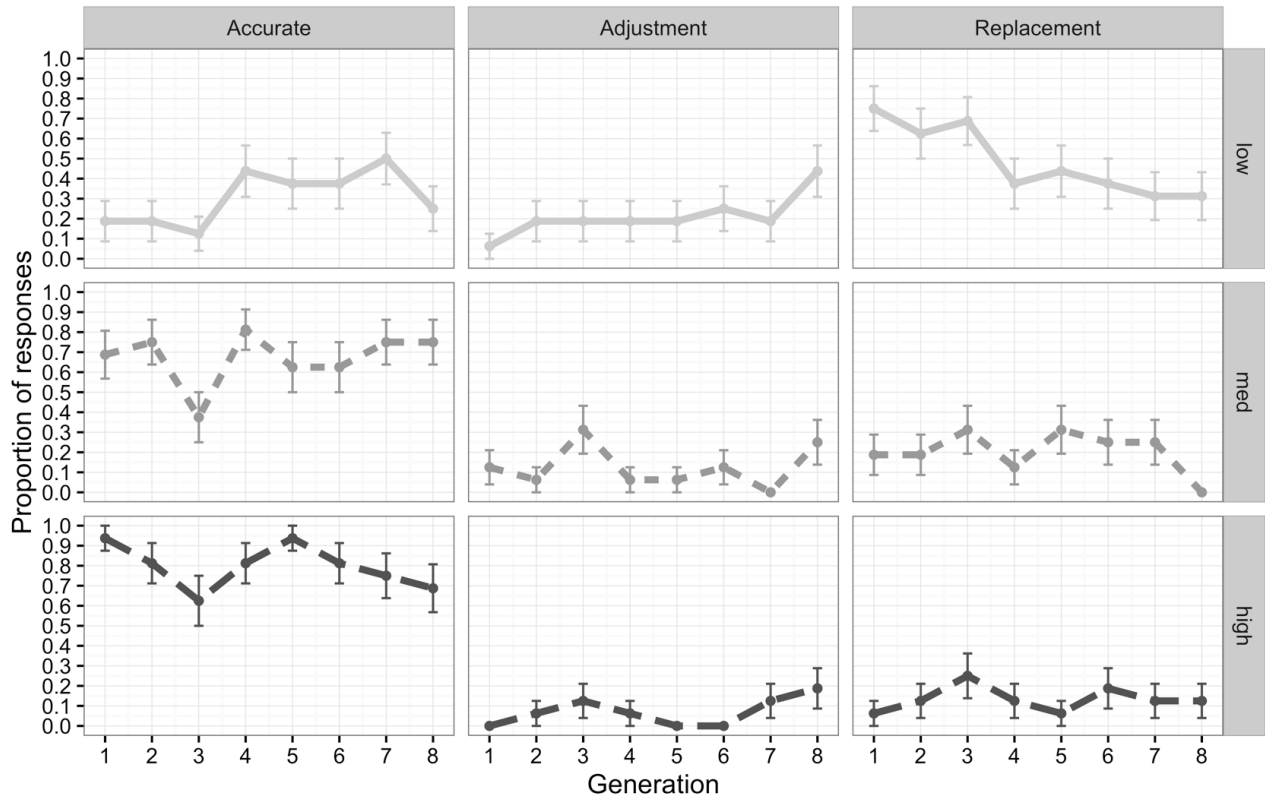
### **Rates of replacement and adjustment**

*Replacements.* For the dependent variable of replacement (whether the response generated during testing was unrelated to the original training word), results from the LRTs revealed that on the full data set, the addition of frequency significantly improved model fit ( $\chi^2(2) = 46.056$ ,  $p < .001$ ), indicating that there were significantly fewer replacements as the frequency of the words increased. Further to this, the addition of generation to the model also improved the fit ( $\chi^2(1) = 7.725$ ,  $p = .005$ ), indicating that over the course of transmission the number of replacements significantly decreased. Additionally, the inclusion of the interaction term frequency x generation revealed a marginally significant improvement to model fit ( $\chi^2(2) = 5.665$ ,  $p = .059$ ), indicating that as the

language was transmitted across generations, there was a marginally significant change in the number of replacements as a result of frequency being varied.

In order to explore this interaction, we then conducted follow up analyses on the subsetting data from each of the frequency conditions. For the low frequency trials, we found a significant improvement to model fit when generation was added to the model ( $\chi^2(1) = 13.033, p < .001$ ), demonstrating that the number of replacements produced by participants was decreasing over the course of transmission. This was not the case for the medium or high frequency conditions, where no significant improvement to model fit was found ( $\chi^2(1) = 0.617, p = .432$  and  $\chi^2(1) = 0.030, p = .862$  respectively), indicating no changes in the number of replacements over the course of the transmission. See [Figure 5.3](#) for results.

*Adjustments.* For the dependent variable of adjustment (whether the response generated during testing was incorrect, but retained similarity to the original training word), results from the LRTs revealed that on the full data set, the addition of frequency significantly improved model fit ( $\chi^2(2) = 11.798, p = .003$ ), indicating that the number of adjustments decreased as frequency increased. Further to this, the addition of generation to the model also improved the fit ( $\chi^2(1) = 4.589, p = .032$ ), indicating that over the course of transmission the number of adjustments significantly decreased. However, the addition of the interaction term frequency x generation did not significantly improve model fit ( $\chi^2(2) = 2.257, p = .324$ ), indicating that as the language was transmitted across



**Figure 5.3.** Mean proportion of responses classified as either accurate, adjustment or replacement in Experiment 7, separated by frequency condition. Error bars show SEM.

generations, there was a marginally significant change in the number of adjustments as a result of frequency being varied.

Next, we performed follow up analyses on the subsetting data from each of the frequency conditions. For the low frequency trials, we found a significant improvement to model fit when generation was added to the model ( $\chi^2(1) = 4.613, p = .031$ ), demonstrating that the number of adjustments produced by participants was increasing over the course of transmission. This was not the case for the medium or high frequency conditions, where no significant improvement to model fit was found ( $\chi^2(1) = 0, p = 1$  and  $\chi^2(1) = 2.110, p = .146$  respectively),

indicating no changes in the number of adjustments over the course of the transmission. See [Figure 5.3](#) for results.

## Discussion

As was the case in [Chapter 4](#)'s results, we found a robust frequency effect during the recall of the languages after participants were trained on words that varied in their frequency of presentation. Interestingly however, the current experiment demonstrated that these effects are reliably maintained over the course of cultural transmission, with high frequency words being recalled more accurately and with less error than words presented less frequently. These results provide evidence for the hypothesis that there is a consistent difference in error between words of high and low frequency that is observable over several generations of learners. Such a finding supports previous reports that high frequency words are evolutionarily more stable than low frequency words, as was the case in [Pagel et al's \(2013\)](#) study, whereby high frequency words maintain higher levels of fidelity. Such a finding bolsters the idea that a word's ancestral roots can be traced back in time, with frequency being available as a reliable predictor of how stable a word form may be in the vocabulary.

Yet it should be noted that our results build a more insightful picture of how lexical items in a language evolve. Whilst previous reports demonstrate a contrast in fidelity as predicted by frequency, we have shown here that over the course of cultural transmission, higher frequency words remain resilient to production error throughout, yet the words presented with low frequency exhibit

an increase in learnability over the course of transmission. This suggests that although a word can be used with low frequency, which increases its likelihood of production error, by transmitting the language culturally, these words can in fact adapt to become more learnable over time. If this is indeed the case, then the finding of Kirby et al (2008), where languages evolve to become more learnable over the course of cultural transmission, then our results may indicate that it is only the low frequency words that may be driving this evolutionary change, with higher frequency words resisting such modification given their relative processing advantages. This would be consistent with Kirby, Dowman and Griffiths (2007) and Lieberman et al. (2007), where high frequency words resist regularization within the language and low frequency words do adopt more regular patterns. Additional support for this claim can be found in Monaghan et al.'s (2014) analysis of systematicity in natural language, where it was reported that a marginally significant relationship exists ( $p = .063$ ) between frequency and the systematicity of the words, where lower frequency words are marginally more systematic. This may provide low frequency words with a processing boost, enabling them to still retain some potential for successful acquisition and use.

Further to this, our experiment also revealed interesting patterns of results for our replacement/adjustment analyses. In Chapter 4, our results demonstrated that low frequency of presentation during training generated more responses during the testing phase that were classified as replacements, when compared with those of higher frequency. We find the same result in our first generation of learners in the present experiment, however whilst the higher frequency words

retain their low levels of replacements throughout the cultural transmission of the language, there is crucially a decrease in the number of replacements for low frequency words. Alongside this, we observe an increase in the number of adjustments being produced for low frequency words. This would suggest that over time the low frequency words are, as demonstrated by our learnability analysis, becoming more learnable, but this increase in learnability may be attributable to the fact that the words are being adjusted more often in later generations, instead of being replaced in the language. This interpretation of the results would be coherent with our previous claim that low frequency words may be adapting to become more learnable, but doing so by adjusting their word forms during processes of cultural transmission, in order to ease the processing constraints placed on them as a consequence of their low frequency of use.

### 5.3 Experiment 8: Word length and iterated learning

#### Method

**Participants.** The experiment was completed by 32 students from Lancaster University (23 females) with a mean age of 22.6 years ( $SD = 6.59$ , range = 18-54). All participants were proficient in English and received £3.50 for participating.

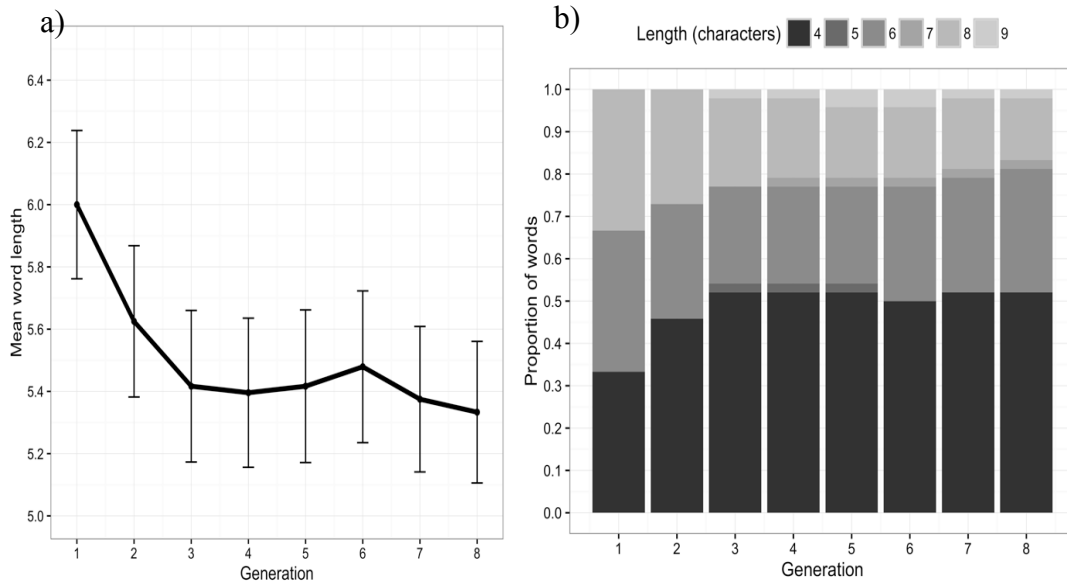
**Materials.** The initial starting languages presented during training to the first generation in each transmission chain were identical to those used in [Chapter 4's length](#) experiment. This meant that the same consonants and vowels were



used for the word forms, as well as the same 12 visual stimuli to represent the meanings (see [p.101](#) for details on their design and [Table 4.1, p.85](#) for examples of the language. Words were generated through a pseudo-random process, ensuring that at the start of the transmission chain, there was no underlying correspondence between the phonology of a word and its combinatorial structure.

**Procedure.** The experiment involved learning and then reproducing an artificial language, where the length of the words was manipulated (see [Chapter 4, p.101](#) for details on how these were manipulated). The procedure was identical to that used in Experiment 7 of the present chapter, but with the manipulation of word length. However, when a word was duplicated during the testing phase, to remove underspecification, a new word was chosen that was of the same length as the originally produced word, for example if the word was 4 letters long then the new word used to replace that word would also be 4 letters long. This ensured that any changes in word length resulting from participants' transmissions were maintained during the process of iterated learning.

**Analysis.** The same analysis was used as in Experiment 7. However, as participant's responses were unconstrained during the testing phase in Experiment 7, i.e. they could provide a response that was composed of any combination of letters and could be of any length, this resulted in much more variation in word length than for Experiment 7. This motivated analyses to test whether words produced by participants varied in length as a result of the transmission of the languages by iterated learning. Results from LRTs where



**Figure 5.4.** a) Mean length of words by generation for Experiment 8. Error bars show SEM. b) Proportional distribution of different word lengths by generation. Both figures are aggregated across all 4 transmission chains.

generation was added to a model predicting word length, revealed a significant decrease in word length as the languages were being transmitted ( $\chi^2(1) = 7.273, p = .007$ ) (see [Figure 5.4a](#)). Additional analyses also revealed more variety in the lengths of the words being used in the languages. Initial starting languages only contained words of length 4, 6, or 8 characters, however as [Figure 5.4b](#) shows, there was also the introduction of 5, 7 and 9 letter long words. In order to classify word length categories, words that were 4-5 letters long were coded as *short*, 6-7 letters long as *medium* and 8-9 letters long as *long*. A multiple regression analysis predicting standard deviation of word length revealed a significant interaction between these word length groups and generation ( $\beta = 0.04, t = 2.699, p = .015$ ), with standard deviations for medium and long words increasing over the course

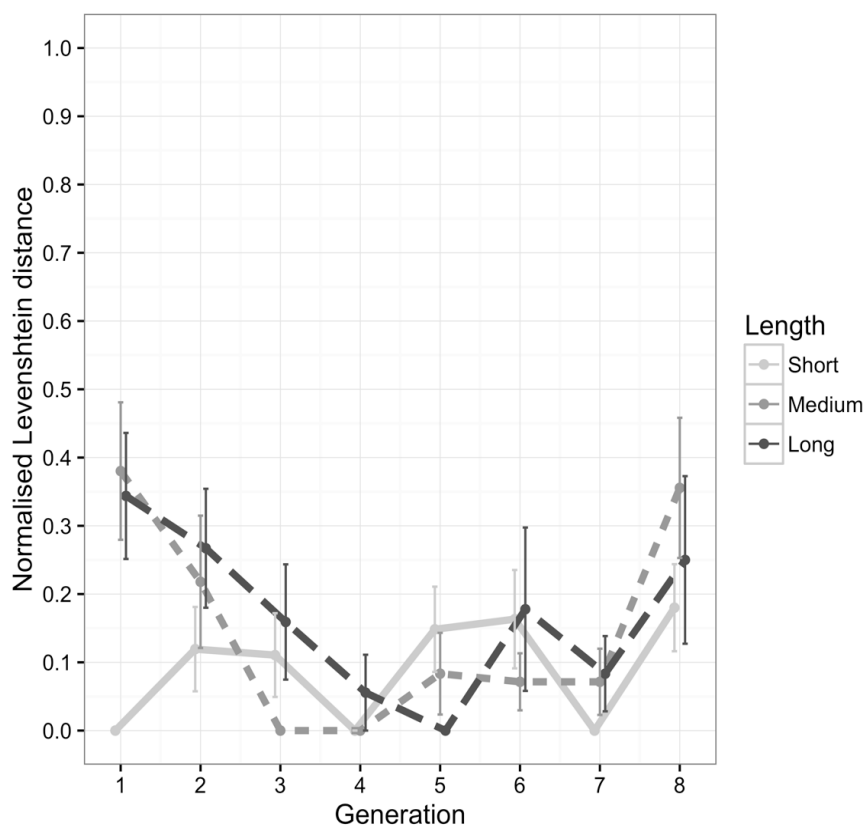
of transmission (both  $p$ 's  $< .05$ ), but no change for short words ( $p = .689$ ). This indicates that over the course of transmission, there was more variation in word length in longer words.

## Results

### Learnability

*Error.* For the dependent variable of normalised Levenshtein distance, results from the LRTs revealed that on the full data set, the addition of length significantly improved model fit ( $\chi^2(2) = 14.151, p < .001$ ), indicating that increasing the length of a word resulted in greater error (estimate = 0.09,  $SE = .03, t = 3.342$ ). However, the addition of generation to the model did not improve model fit ( $\chi^2(1) = 0.367, p = .545$ ), indicating that over the course of transmission there were no changes in the error made by participants. We also found no significant improvement in model fit with the addition of the interaction term length x generation ( $\chi^2(3) = 3.965, p = .265$ ), indicating that as the language was transmitted across generations, there was no significant changes in error as a result of length being varied. This was highlighted further when we analysed the subsetting data to assess the effect of generation in each of the length conditions, all of which showed no significant improvement to model fit when generation was added as a predictor (short:  $\chi^2(1) = 1.385, p = .239$ ; medium:  $\chi^2(1) = 0.250, p = .617$  and long:  $\chi^2(1) = 1.607, p = .205$ ). See [Figure 5.5](#) for results.

*Accuracy.* Results from the LRTs revealed that, when assessing the number of accurate responses, the addition of length significantly improved



**Figure 5.5.** Mean error of production by generation in the transmission chain for Experiment 8. Values were calculated from the normalised Levenshtein distance between the training and production words for each of the length classification. Error bars show SEM.

model fit ( $\chi^2(2) = 13.91, p < .001$ ), indicating that the number of accurate productions made by participants during the testing phase decreased as the word length increased. However, the addition of generation to the model did not improve the fit ( $\chi^2(1) = 0.318, p = .573$ ). However, the addition of the interaction term length x generation did significantly improve the model fit ( $\chi^2(3) = 9.321, p = .025$ ).

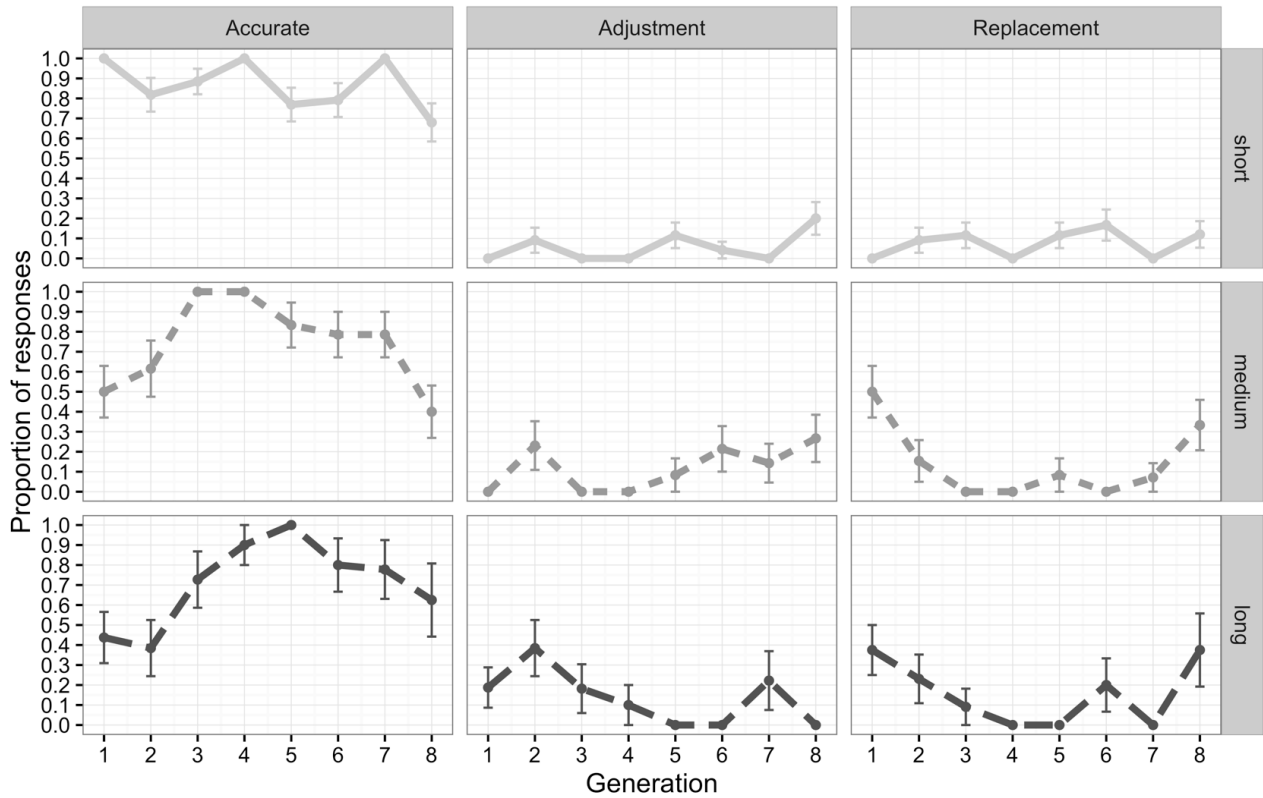
In follow up analyses to explain this interaction, there was a marginally significant improvement to model fit for the short words when generation was

added ( $\chi^2(1) = 3.063, p = .080$ ), suggesting that accuracy was decreasing marginally over the course of transmission. This was not the case for the medium length words, where no significant improvement to model fit was found ( $\chi^2(1) = 0.048, p = .827$ ), indicating that the number of accurate responses remained stable even when the language was being transmitted across generations. However, for the long words we observed an improvement to model fit with the addition of generation ( $\chi^2(1) = 6.025, p = .014$ ), indicating that accuracy improved over the course of transmission. See [Figure 5.6](#) for results.

Note that this significant increase in learnability for our accuracy measurement in long words was not observed for our error measurement (normalised Levenshtein Distance). This may be interpreted as participants attending to a particular long word in the language in order to ensure accurate learning of that word, hence the increase in accuracy. Whilst this would also decrease error, attending to a particular long word may incur a cost for the learning of other long words, as cognitive resources may now be limited, thus this may cancel out any effects of increased accuracy.

### **Rates of replacement and adjustment**

*Replacements.* For the dependent variable of replacement (whether the response generated during testing was unrelated to the original training word), results from the LRTs revealed that on the full data set, the addition of length significantly improved model fit ( $\chi^2(2) = 10.633, p = .005$ ), indicating that there were fewer replacements as word length decreased. However, the addition of generation to the model did not improve the fit ( $\chi^2(1) = 0.961, p = .327$ ),



**Figure 5.6.** Mean proportion of responses classified as either accurate, adjustment or replacement in Experiment 8, separated by length classification. Error bars show SEM.

indicating that over the course of transmission the number of replacements did not significantly change. Similarly, the addition of the interaction term length  $\times$  generation also revealed no significant improvement to model fit ( $\chi^2(3) = 2.788, p = .425$ ). This was also reflected by subsequent analyses on each of the length conditions, where none of the models improved with the addition of generation (short:  $\chi^2(1) = 0.484, p = .487$ ; medium:  $\chi^2(1) = 1.561, p = .212$  and long:  $\chi^2(1) = 0.422, p = .516$ ). See Figure 5.6 for results.

*Adjustments.* For the dependent variable of adjustment (whether the response generated during testing was incorrect, but retained similarity to the

original training word), results from the LRTs revealed that on the full data set, the addition of length significantly improved model fit ( $\chi^2(2) = 9.509, p = .009$ ), indicating that there was a significant increase in the number of adjustments produced as the length of the word increased. However, the addition of generation to the model did not improve the model fit, ( $\chi^2(1) = 1.654, p = .198$ ), indicating that over the course of transmission the number of adjustments did not change overall. Importantly, the addition of the interaction term length x generation did significantly improve model fit ( $\chi^2(3) = 11.578, p = .009$ ).

Next, we performed follow up analyses on the subsetting data from each of the length conditions. For the short words, we found a marginally significant improvement to model fit when generation was added to the model, ( $\chi^2(1) = 3.634, p = .057$ ), demonstrating that the number of adjustments increased marginally over the course of transmission. There was a significant improvement to model fit for the medium length words ( $\chi^2(1) = 4.139, p = .042$ ), indicating that the number of adjustments increased over the course of transmission. For the long words, there was also a significant improvement to model fit ( $\chi^2(1) = 3.889, p = .049$ ), however this change was a reduction in adjustments over the course of transmission. See [Figure 5.6](#) for results.

## Discussion

Again we have found evidence of a length effect in our results, where shorter words were recalled with less error than longer words, building on the findings reported in Chapter 4, but here we demonstrate this effect also remains even

when a language has been culturally transmitted across several generations of learners. This was also the case for the number of replacements being produced by the learners, with longer words being replaced more often than shorter words. These findings demonstrate that [Monaghan's \(2014\)](#) results can be generated in a laboratory analogue, whereby the rate at which a word undergoes dramatic change in a language can be predicted by the length of the word form used to represent its meaning. However, neither of our error or replacement measurements indicate that these values change as a result of the cultural transmission of the language.

Yet there is evidence of change occurring as a result of the language being transmitted. We found that shorter words may be decreasing slightly in overall accurate responses, whilst longer words show the inverse, with accuracy increasing over time. Our analyses revealed that there was a significant trend for word lengths to be reduced over the course of transmission, this suggests that participants would shorten words in the language. This could explain why shorter words were seemingly being recalled less accurately, as more words were becoming shorter this would increase confusability among words, in contrast to longer words which would emerge as distinctive forms in the language, increasing their likelihood to be recalled accurately. Moreover, our results suggest that adjustments were increasing over the course of transmission for shorter words, again this could be explained by the fact that there was increased competition between shorter words for distinctiveness in the language, this could lead to small errors being made during production.



Word length is a particularly complex feature of language, with its function extending much further than one would initially expect. Given that word length is not only intimately linked with frequency (Zipf, 1949), but also information content (Piantadosi et al., 2011) and conceptual complexity (Lewis & Frank, 2015; 2016), then it would be of particular interest to consider how these other factors may enhance the findings reported here. Would words that carry low information content and low conceptual complexity also exhibit resistance to change over the course of cultural transmission? Such questions would be of considerable interest, but at the moment remain a topic for future research.

## 5.4 Experiment 9: AoA and iterated learning

### Method

**Participants.** The experiment was completed by 32 students from Lancaster University (22 females) with a mean age of 25.6 years ( $SD = 5.03$ , range = 18-38). All participants were proficient in English and received £3.50 for participating.

**Materials.** The initial starting languages presented during training to the first generation in each transmission chain were identical to those used in Chapter 4's AoA Experiment. This meant that the same consonants and vowels were used for the word forms, as well as the same 12 visual stimuli to represent the meanings (see p.115 for details on their design and Table 4.1 for examples of the language. Words were generated through a pseudo-random process, ensuring that

at the start of the transmission chain, there was no underlying correspondence between the phonology of a word and its combinatorial structure.

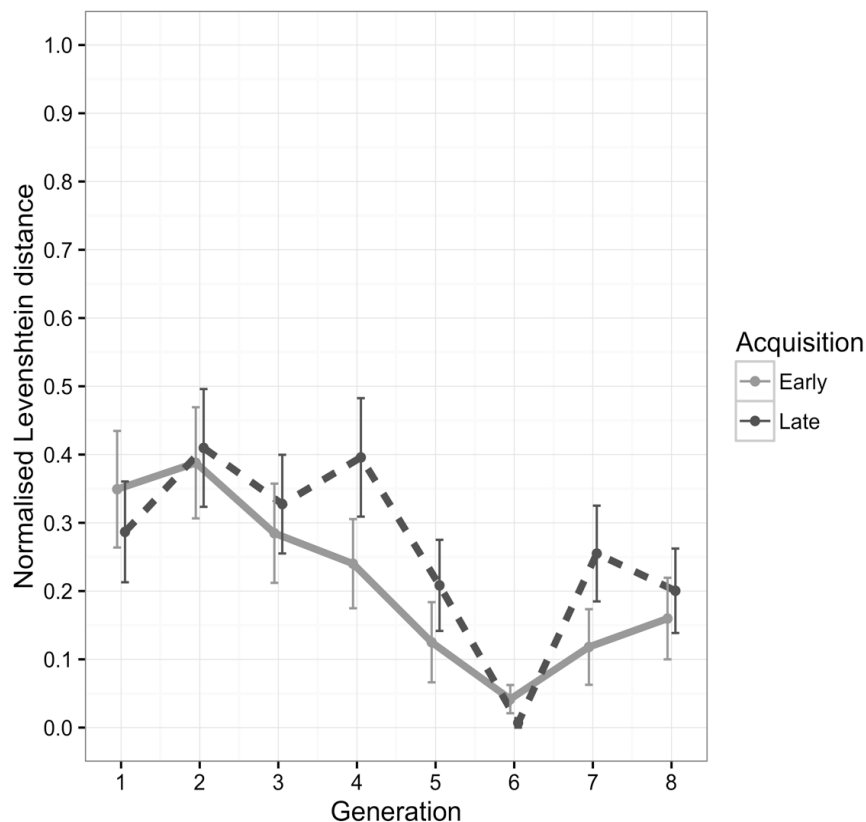
**Procedure.** The experiment involved learning and then reproducing an artificial language, where the AoA of the words were manipulated (see [Chapter 4, p.115](#) for details on how these were manipulated). The procedure was identical to that used in Experiment 7 of the present chapter, but with the manipulation of AoA.

**Analysis.** The same analysis was used as in Experiment 7.

## Results

### Learnability

*Error.* For the dependent variable of normalised Levenshtein distance, results from the LRTs revealed that on the full data set, the addition of AoA did not significantly improve model fit ( $\chi^2(1) = 1.876, p = .171$ ), indicating that there was no significant difference in the amount of error participants made during testing for the early or late acquired words. However, the addition of generation to the model did improve model fit ( $\chi^2(1) = 22.772, p < .001$ ), indicating that over the course of transmission error significantly decreased. We found no significant improvement in model fit with the addition of the interaction term AoA x generation ( $\chi^2(1) = 0.367, p = .425$ ), indicating that as the language was transmitted across generations, there was no differences in error for the early and late acquired words.



**Figure 5.7.** Mean error of production by generation in the transmission chain for Experiment 9. Values were calculated from the normalised Levenshtein distance between the training and production words for each of the two AoA conditions. Error bars show SEM.

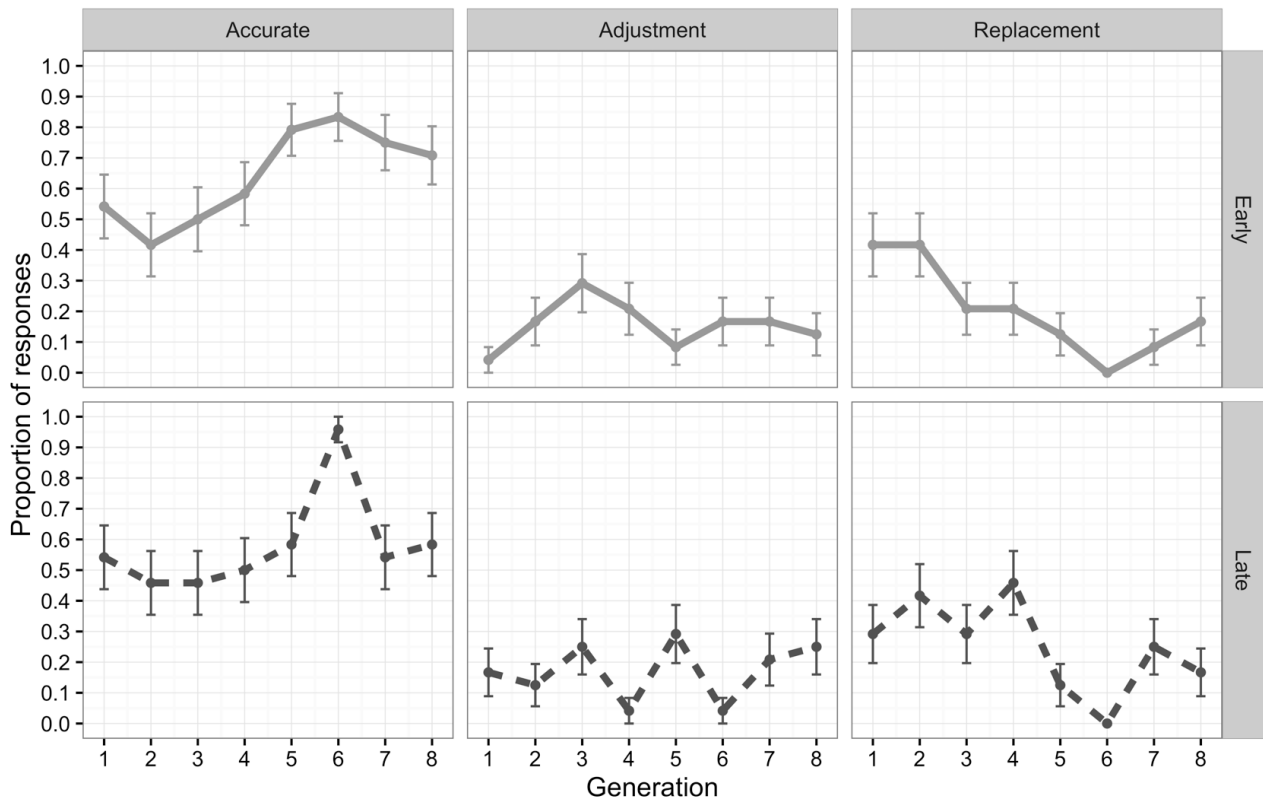
This was highlighted further when we analysed the subsetting data to assess the effect of generation in each of the AoA conditions, with both early and late acquired words showing significant improvements to model fit when generation was added as a predictor (early:  $\chi^2(1) = 16.931, p < .001$  and late:  $\chi^2(1) = 7.356, p = .007$ ), thus showing that although both AoA conditions demonstrated decreases in error over the course of transmission, these did not differ from each other in the amount of error participants produced. See [Figure 5.7](#) for results.

*Accuracy.* Similar results were found from the LRTs, when assessing the number of accurate responses produced, with the addition of AoA yielding no significant improvement to model fit ( $\chi^2(1) = 1.593, p = .207$ ), indicating that the number of accurate productions made by participants during the testing phase did not vary for early and late acquired words. However, the addition of generation to the model did improve the fit ( $\chi^2(1) = 12.286, p < .001$ ), indicating that accuracy increased over the course of transmission. Although there was still no significant improvement to model fit with the addition of the interaction term AoA x generation ( $\chi^2(2) = 2.858, p = .240$ ), indicating that accuracy did not differ by AoA over the course of transmission.

In follow up analyses however, for the early acquired words the addition of generation to the model significantly improved the fit ( $\chi^2(1) = 10.750, p = .001$ ), indicating a significant increase in accuracy over the course of transmission. For the late acquired words however, this only resulted in a marginal improvement to model fit ( $\chi^2(1) = 3.091, p = .079$ ), demonstrating that accuracy only improved marginally over the course of transmission. See [Figure 5.8](#) for results.

### **Rates of replacement and adjustment**

*Replacements.* For the dependent variable of replacement (whether the response generated during testing was unrelated to the original training word), results from the LRTs revealed that on the full data set, the addition of AoA did not significantly improve model fit ( $\chi^2(1) = 1.210, p = .271$ ), indicating that there were no differences in the number of replacements produced by participants



**Figure 5.8.** Mean proportion of responses classified as either accurate, adjustment or replacement in Experiment 9, separated by AoA condition. Error bars show SEM.

based on the AoA of the words. However, the addition of generation to the model did improve the fit ( $\chi^2(1) = 20.833, p < .001$ ), indicating that over the course of transmission the number of replacements significantly decreased. The addition of the interaction term AoA x generation revealed no significant improvement to model fit ( $\chi^2(2) = 2.740, p = .254$ ), indicating that there was no difference between the number of replacements generated by early and late acquired words over the course of transmission. This was also reflected by subsequent analyses on each of the AoA conditions, where both of the models improved with the

addition of generation (early:  $\chi^2(1) = 15.744, p < .001$  and late:  $\chi^2(1) = 6.579, p = .010$ ). See [Figure 5.8](#) for results.

*Adjustments.* For the dependent variable of adjustment (whether the response generated during testing was incorrect, but retained similarity to the original training word), results from the LRTs revealed that on the full data set, the addition of AoA did not significantly improve model fit ( $\chi^2(1) = 0.171, p = .679$ ), indicating that there was no significant difference in the number of adjustments produced as the AoA of the word varied. Similarly, the addition of generation to the model did not improve the model fit ( $\chi^2(1) = 0.261, p = .609$ ), nor did the addition of the interaction term AoA x generation ( $\chi^2(3) = 0.571, p = .903$ ) indicating that over the course of transmission the number of adjustments did not change, nor did they differ based on the AoA of the words during this transmission. This was supported further by subsequent analyses on the early and late acquired words, with neither of the conditions improving their model fits with the addition of generation (early:  $\chi^2(1) = 0.008, p = .931$  and late:  $\chi^2(1) = 0.393, p = .531$ ), indicating that the number of adjustments produced did not change for either of the AoA conditions during transmission. See [Figure 5.8](#) for results.

## Discussion

We aimed in this experiment to explore the role of AoA on lexical stability throughout the course of the cultural transmission of language. Similar to our findings reported in [Chapter 4](#), we found no clear difference in early and late

acquired words error based on our normalised Levenshtein distance measurement, with both sets of words decreasing in error over the course of transmission. However, we did find a difference, although relatively modest, in the way that accuracy changes: early acquired words showed an increase in accuracy as the language was transmitted across generations, whilst this was observed to be a much weaker trend for late acquired words. This would suggest that early acquired words are in some respect, exhibiting an increase in their fidelity as a result of the language being transmitted culturally. If there is indeed a distinction based on accuracy and AoA, then this would suggest that early acquired words contribute to lexical stability as suggested by [Monaghan \(2014\)](#) and would importantly contradict claims that the point of change during language evolution has its locus on the child learner ([Bickerton, 1990](#)) – lexical change is instead the province of words acquired later in language development.

Yet the present experiment did not find any significant differences in the number of replacements produced by the learners. Whilst we predicted late acquired words to be replaced more often than those acquired earlier, the two word sets demonstrated a trend to decrease the number of replacements over the course of transmission, but with no clear difference between the two word types. This finding does not provide direct evidence for [Monaghan's \(2014\)](#) finding, where there is such a distinction, however it also does not provide valid support for the claim that change occurs as a consequence of the acquisition process, as highlighted by the fact that late acquired words are also not being observed to be replaced at a lower rate.

Whilst the results presented here may only reveal a very subtle effect for AoA, one could raise concerns over the relationship between properties of words acquired in natural language and those used in the artificial language here. If we consider the importance of sound-symbolism to the learner and its underlying incorporation within the vocabulary (Monaghan et al., 2014), in addition to the benefits it provides to the learner during the acquisition process (Chapter 2), then this may play an important role in the way that early acquired words could be conserved in the vocabulary and resist change over time. Thus, capturing the rich and complex properties of words in the stimuli of such a paradigm as used here, could allow for differences between the AoA of words to be teased apart with more validity.

## 5.5 General discussion

The application of the iterated learning model to demonstrate how a language changes as a result of cultural transmission has previously aimed to explore questions regarding the emergence of certain properties of language (Kirby et al., 2008). Here, we aimed to explore not only what causes change in a language, but also the causes of resistance to change, leading to certain properties being conserved in the language. In Chapter 4, it was demonstrated that psycholinguistic properties of lexical items can lead to distinctions between the magnitude of production errors, in this chapter we have demonstrated that these distinctions are retained over the course of iterated learning. Words that are higher in frequency and shorter in length are retained with higher fidelity across



generations of learners, although the distinction is much more subtle for AoA. Interestingly, the words that were presented with lower frequency and longer length, were observed to adapt to the learner, suggesting that the process of iterated learning may serve to increase learnability of items that are processed with more difficulty, whilst words processed relatively easily do not require such adaptation and therefore do not undergo such change. This would support previous findings that suggest that processes of change are located where processing limitations are found ([Kirby, 2001](#); [Kirby et al., 2007](#); [Lieberman et al., 2007](#); [Monaghan, 2014](#)), and this change is a direct result of the language being transmitted culturally.

Yet there are still limitations to the present experiment. The use of only four transmission chains does not yield particularly strong effects. If we consider the power required for reliable effects to be observed in traditional psycholinguistic experiments, then the comparatively small number of participants allocated to each generation within a transmission chain must be increased. This may explain why there was only a weak effect for the AoA experiment, which would benefit from a larger sample size to uncover effects that are known to be more subtle than effects of, for instance, frequency and length (see [Monaghan, 2014](#)). Furthermore, the present experiments have indeed only directly addressed questions relating to the evolution of nouns in a language. Given the range of different parts of speech within a language, there would be reason to question how generalisable the findings presented here are to words from other grammatical categories. This would be of particular interest for future

work that could adapt the current experimental design in order to investigate learning of verbs ([Monaghan, Mattock, Davies, & Smith, 2015](#)) or even words that are from a closed word class. This would be especially interesting considering previous studies of language change, which have consistently demonstrated that part-of-speech is the greatest contributor in models predicting variance in the likelihood of change occurring, above any other psycholinguistic variable ([Pagel et al., 2007](#); [Tadmor, 2009](#)).

Importantly however, we have here a paradigm for testing predictions relating processes of cognition to processes of language change, providing a new perspective to an area that has long focused primarily on the social and geographic explanations of change in language ([Labov, 2001](#); [Nettle, 1999](#); [Thomason & Kaufman, 1988](#)). Moreover, introducing such psycholinguistic properties within the design of iterated language learning experiments in the laboratory, allows for the rich dynamics of language to be incorporated into the tests and thus enable investigations into the cognitive factors driving languages to change and evolve.

# Chapter 6

## Conclusions

---

### 6.1 Aims of thesis

At the core of this thesis is the overarching hypothesis that processes of acquisition influence the evolution of language. In order to establish a direct link between acquisition and evolution, this thesis presents evidence which demonstrates how cognitive factors shape the structure of the developing vocabulary, then relates these factors to the stability of words in the language over the course of language evolution and change. We argue that properties such as non-arbitrariness are present early on in language development, because they benefit the learner and therefore need are preserved over time. In contrast, properties such as arbitrariness are prevalent in later language, because they are more suited to meet the needs of the language user, therefore changes in word forms are not problematic as there is no underlying relationship tying form to

meaning, and may in fact be adapting to conform to patterns of systematicity present in the language.

This thesis has aimed to present a view of human language which is dynamic, incorporating properties that can aid its acquisition early on in life, such as non-arbitrariness, but then adapting to address the needs of a more mature user of the language, where communicative pressures shape the nature of form-meaning mappings learnt later on in life.

The theoretical motivations for exploring this aim empirically, derive from previously reported findings from computational models of word learning, which reveal an advantage for non-arbitrariness when the vocabulary size is small, which is the case during early stages of acquisition, yet arbitrariness provides a more optimal system when the vocabulary size increases, which is normally the case for more mature language users ([Gasser, 2004](#)). However, such results have not been demonstrated behaviourally, therefore the thesis aimed to establish a novel methodology to test such predictions, where the relative benefits of arbitrariness and non-arbitrariness can be examined in the context of a developing vocabulary (see [Experiments 1-3](#)).

Developing on from this first aim, the thesis also presents an evolutionary perspective on how the vocabulary might adapt over time, as a consequence of the language being culturally transmitted over generations of learners and users. Here, a second main aim was to demonstrate that during this transmission process, certain features in the vocabulary are preserved from dramatic changes, whilst also assessing why other features are more vulnerable to these changes.

The primary focus of this aim was to provide explanations as to what is driving such a dissociation in the rates of linguistic change and evolution.

Once again, the current literature surrounding this topic has provided some initial insights on the possible mechanisms driving differences in rates of change. Results obtained from phylogenetic studies have indicated that psycholinguistic properties, such as word frequency, length and AoA, can be used to reliably predict how rapidly a word undergoes replacement by a new non-related cognate form ([Monaghan, 2014](#); [Pagel et al, 2007](#)). However, the robustness of such findings is somewhat limited, with results obtained from corpora consisting of a relatively small number of words, therefore providing subsequent evidence is crucial to support these claims. Recent advances in the way language evolution can be studied under laboratory conditions, has provided researchers with the means to test such predictions (see [Tamariz, 2017](#) for review), therefore we present experimental evidence to support these previous findings, whereby we can observe the rate and types of changes that words undergo during transmission (see [Experiments 4-9](#)).

Empirical evidence has been sought within this thesis to investigate the main hypotheses presented in this section. This evidence will be summarised in the next section, with the implications of the findings also being discussed.

## 6.2 Summary of findings

Chapters 2 and 3 were primarily focused on examining the role of sound-symbolism and arbitrariness in language learning, examining how the nature of

mappings between form and meaning can influence the learning of words in different ways. Although there has been copious research which demonstrates an advantage for sound-symbolism when distinguishing between categories of meanings (Farmer et al., 2006; Luypan & Casasanto, 2015; Mongahan et al, 2012), these chapters present a series of experiments which reveal when sound-symbolism and arbitrariness may influence the learning of individuated meanings in a language by manipulating vocabulary size.

Experiment 1 presented a study where a language comprised mappings that were either congruent or incongruent with previously established sound-symbolic associations between linguistic sounds and visual shapes. Whilst such a design has been used before to measure performance during cross-situational learning (Monaghan et al, 2012), there has thus far been no behavioural investigation examining how performance varies when the vocabulary size is changed to reflect different stages of language development. This experiment aimed to do just that.

The results revealed that sound-symbolism facilitated learning of categorical distinctions in a large vocabulary size, but not individuated meanings in the language. However, an import finding was observed in the small vocabulary size, where sound-symbolism was reported to benefit the acquisition of individuated meanings. This has implications for theoretical accounts of sound-symbolism and language learning, because it provides evidence which emphasises the importance of non-arbitrariness for the learning of individual items in the vocabulary. However, this effect dissipates as the vocabulary size

increases, where the facilitatory role of sound-symbolism shifts towards aiding the learning of categorical distinctions present in the vocabulary.

Experiments 2 and 3 examined the way that learning was affected when languages were designed to be either fully sound-symbolic (Experiment 2) or where none of the mappings displayed a relationship between form and meaning (Experiment 3). Both of the experiments included the additional manipulation of vocabulary size, thus this would provide a more comprehensive examination of the way sound-symbolism and arbitrariness function within small and large vocabulary sizes, providing an assessment of how languages may function when only one system is used to map form to meaning (similar to [Monaghan et al, 2011](#)).

Results from Experiment 2 indicated that when the language was fully sound-symbolic, performance during categorical learning was significantly greater than when learning individual mappings. Although this was a trend observed in all vocabulary size conditions, the greatest benefits were observed in the large vocabulary size. This indicates that an exclusively sound-symbolic language would be beneficial to the learner when distinguishing across distinct categories. Such a finding is in line with corpus analyses of the vocabulary, where systematicity exists in the phonology of words, which can be used to discriminate between certain categories of words ([Kelly, 1992](#); [Monaghan et al, 2007](#); [2012](#)). No significant effects of vocabulary size were found for individual word learning, meaning the evidence in support of a learning advantage for sound-symbolism in a small vocabulary size was only reported in Experiment 1,

when there was variation in the types of form-meaning mappings present in the language.

Experiment 3 provided insights into how a fully arbitrary vocabulary affects learning. Here, we presented a novel methodology for quantifying the extent to which different word forms may exhibit a systematic relationship to different angular and rounded shapes. Whilst this methodology was used to confirm previous assertions about sound-symbolic mappings, it also offered insights into when words display no systematic relationship between form and meaning. Previous research has largely examined sound-symbolism effects by using experimental designs that use congruent and incongruent sound-symbolic mappings (e.g. [Monaghan et al., 2012](#)). As incongruent mappings incorporate aspects of non-arbitrariness as they are still relatively systematic, this may not necessarily be testing the effects of arbitrariness in language, where no relationship exists. By developing an experimental design where the mappings are known to hold no underlying relationship, we can reliably test arbitrariness in the laboratory, just as we can with sound-symbolism.

The results from this experiment revealed another trend for better performance during categorical learning trials across all vocabulary sizes, when compared with individual word learning trials. Whilst this trend was observed in Experiment 2, important differences can be observed between the two experimental results. By comparing the results from the two experiments, it was revealed that there was significantly higher accuracy when the vocabulary was sound-symbolic ([Experiment 2](#)) in contrast to when it was fully arbitrary



([Experiment 3](#)), suggesting that the benefit for categorical learning was ultimately more prominent when there was systematicity in the language.

Finally, the results of Experiment 3 also revealed a subtle effect for learning individual form-meaning mappings in a large arbitrary vocabulary, where over the course of learning participants' accuracy improved. This contrasts to the results from the small arbitrary vocabulary, where there was no evidence that participants' accuracy increased, indicating that learning only occurred for the large arbitrary vocabulary.

In light of the distinctions reported between the developing vocabulary, Chapters 4 and 5 aimed to establish a direct link between acquisition and evolution, by exploring the role of psycholinguistic properties in the way that language changes over time. An artificial language learning paradigm was developed which assessed how variations in word frequency, length and AoA (properties known to vary over the course of vocabulary development) might result in variance in the rate of linguistic change and evolution. Moreover, we also generated a novel analysis to quantify the types of change that a word might undergo, making important distinctions between adjustments (where the change is small) and replacements (where there is a substantial change). This allowed for finer grained analyses of our data, which expands on the previous cladistics studies, where analyses was focused on predicting rates of replacement and less so on adjustments (e.g. in [Monaghan, 2014](#); [Pagel et al, 2007](#), [Vejdemo & Hörberg, 2016](#)).

Chapter 4 presented experimental data from a series of experiments that assessed how a single generation of learner may introduce changes in a language. The results demonstrated that when learning words which varied in frequency (Experiment 4), the number of accurate recalls was higher for words presented frequently, whereas low frequency words had more replacements. There were no significant differences for adjustments. Whilst for word length (Experiment 5), shorter words were recalled more accurately than longer words, with no differences between adjustments and replacements in the language. AoA (Experiment 6) revealed no significant differences in either accuracy, adjustments or replacements.

Chapter 5 then went on to explore how the process of cultural transmission across several generations of learners may provide insights into how word frequency, length and AoA might influence language evolution. In Chapter 4, the methodology provided a means to examine the influence of psycholinguistic properties on learnability within an individual learner. However, in order to gain a much more comprehensive understanding of how these properties might influence language evolution, where languages change over many generations of learners, we needed to incorporate an additional dynamic – cultural transmission. Applying the same basic methodology and analyses as used in Chapter 4, the next series of experiments then passed on the output language produced by one generation of learner onto the next generation of learners. Thus, we were able to model the evolution of language, and crucially, examine how changes consequent on learning, not just learnability.

The results demonstrated once more that higher frequency and shorter length are reliable predictors for lexical stability, with consistently high accuracy over the course of transmission ([Experiments 7 & 8](#)). However, for AoA there was only a subtle difference between early and late acquired words, with both sets of words following similar patterns in accuracy, adjustment and replacement rates ([Experiment 9](#)).

Interestingly, there was evidence to suggest that low frequency words adapted over the course of cultural transmission in a way that was not observed for the high frequency words. As predicted by the results of [Pagel et al's \(2007\)](#) study, the low frequency words had the highest rates of replacements in the language, however over the course of transmission, the number of replacements decreased, with the number of adjustments increasing too, resulting in greater accuracy for the low frequency words. These results were interpreted as evidence for low frequency words adapting to the needs of the learner, whereby they were initially difficult to learn, resulting in high likelihood of replacement, but as the language evolved, these words may have undergone changes that reflected a need for greater learnability, gradually changing by adjustment. Such a process has been reported in computational modelling ([Kirby, 2001](#)) and experimentally ([Kirby et al, 2008](#)), but in the work presented here, we provide empirical evidence that tests a non-uniform frequency distribution, offering insights into how the vocabulary changes differentially as a result of psycholinguistic factors.

Within the empirical evidence presented in this thesis, several important findings have been made which have important implications for our

understanding of the nature of human language. In particular, the focus has been on highlighting the role of psycholinguistic factors in the way that the vocabulary develops and then evolves. Yet there are still several lines of enquiry and considerations that should be taken into account when advancing further on this topic of investigation. The next section will address such issues.

### 6.3 Limitations and future directions

The experiments presented in [Chapters 2 and 3](#) focused on the role of sound-symbolism and arbitrariness in language learning, highlighting the important roles they play during different stages of vocabulary development. Yet there remains to be a fully comprehensive examination of the way word forms are used naturally to aid language learning. Given the multi-modal nature of language and the variety of cues we utilise in speech, especially when that speech is directed towards infants, one outstanding question is how such factors are incorporated to enhance potential non-arbitrariness in the language. To what extent do we incorporate additional information when using speech to enhance the non-arbitrariness of words? Is it the case that we utilise multiple cues, such as prosody, gesture or distributional information, which are known to greatly assist in referential communication ([Fernald, 1991](#); [Monaghan, 2017](#); [Nygaard et al., 2009b](#)) in order to facilitate learning in infancy? And if these cues are used by the infant to convey meaning ([Hupp & Jungers, 2013](#)), how might this reinforce their learning of words in the vocabulary? Such questions would provide a more naturalistic understanding of sound-symbolism, whereby arbitrariness at the

phoneme level can be minimized by incorporating a range of non-arbitrary properties during production.

In order to provide more robust and ecologically valid evidence to support the findings on sound-symbolism presented in this thesis, one way would be to adapt the experimental paradigm in order to test infants sensitivity to the individual/categorical learning distinctions. Whilst in our experiments we find strong evidence to suggest sound-symbolism is advantageous for word learning in a small vocabulary, this was tested in adult participants, who have already acquired at least one language prior to the experiment. If these results are directly applicable to first language acquisition, then one should be able to test such hypotheses and find similar results in a younger population of learners.

Whilst, previous research has made claims about sound-symbolism and word learning using infant participants ([Imai et al., 2008](#); [Kantarzis et al., 2011](#); [Ozturk et al., 2013](#)), such studies tend to adopt paradigms which only test categorical learning, not individual word learning (the importance of which is highlighted in [Chapter 2](#) and by [Monaghan et al., 2012](#)). Therefore, although there is some support from infancy research on how sound-symbolism may benefit word learning, the evidence is only indirect so far. Thus, the experimental paradigms used in [Chapters 2](#) and [3](#), could potentially be adapted for younger learners, given the widespread use of cross-situational learning paradigms in infancy research (e.g. [Smith & Yu, 2008](#)).

Within [Chapters 4](#) and [5](#), novel experimental research was used to verify results about language change, which were originally derived from corpus based

approaches. This combination of different approaches to examine similar hypotheses has clear advantages, not least the robustness and reliability of the overall evidence. Yet, there is great potential to utilise other approaches to further enhance the findings accrued so far, both in this thesis and the wider literature. Of particular interest and importance, would be to consider a much wider scope of linguistic change than only confined to vocabulary adaptation. It has long been known that change does not operate solely on the vocabulary of a language, instead other structures in languages, such as phonology and syntax, are also subjected to pressures which result in change and evolution over time. Assessing the role of cognitive factors on these aspects of language, will provide researchers with a greater understanding of how processes, such as acquisition and use, contribute to the way languages change and evolve. Such findings could even shed light on language extinction, with a detailed explanation of which features of language are vulnerable to change and which might be conserved in the future.

Yet, it is important to highlight some key concerns when taking an experimental approach to the study of language change, and even in psycholinguistics more broadly. Although the results presented in the chapters of this thesis provide empirical insights and support for the hypotheses we aimed to investigate, one should acknowledge that these results may not necessarily be derived from sufficiently powered experiments (see [Lakens, 2014](#) for detailed discussion on the importance of statistical power). Whilst every attempt was made to ensure the experiments were carefully designed and conducted to high

standards of scientific rigor, constraints on sample sizes and testing time may limit the validity of our evidence for those hypotheses.

For instance, in [Experiment 4](#), a sample of 21 participants yielded statistically significant support for our hypothesis that the frequency of a word influences the learnability of the word. Our results showed that low frequency words had higher rates of error than high frequency words, using linear mixed-effects models. Yet, when we calculate the statistical power of the results, there is reason for moderate caution. Following [Westfall, Kenny and Judd \(2014\)](#), we conducted a power analysis from the results, finding that  $d = 0.635$ , which could be interpreted as a moderately powered experiment. However, in [Experiment 7](#), where the same methodology was used, but within an iterated learning paradigm, the power analysis reveals much more alarming results, with  $d = 0.126$ . Having such a low powered experiment does not necessarily invalidate the results, but it should raise concern when interpreting the results as support for the hypotheses.

A primary reason for why [Experiment 7](#) (as well as other iterated learning experimental studies in this thesis, and in the wider literature) are so underpowered is the fact that only 4 chains of learners are used. When comparing the sample size used in [Experiment 4](#), where there was moderate statistical power, this would suggest that we would require at least 21 chains of learners, to ensure the same moderate power for our iterated learning experiments. Adopting such an approach would require 21 participants for each of the 8 generations of learners, requiring 168 participants for each iterated learning experiment. Whilst this would have required much greater testing resources, it does highlight an

important methodological consideration for anybody using the iterated learning paradigm in the laboratory, and an additional consideration when interpreting the results from previous work using the paradigm.

Another point of discussion regarding the iterated learning paradigm is the way the generation factor is treated within the analysis. In the studies presented within [Chapter 5](#), generation was taken to be a within-subjects factor, then a regression was used to examine how the factor might predict changes in how the language changed. Yet, the experimental paradigm is structured in a way where new participants are recruited at each new generation. Using such a design therefore, may mean the generation factor should not necessarily be considered as within-subjects, as we are not measuring the same participant's responses at each generation. Instead, one could argue that the design is in fact using a between-subjects factor, where each individual generation should be treated as an independent group. In order to analyse the data suitably, [Winter and Wieling \(2016\)](#) recommends using mixed-effects models where the random effects structure comprises intercepts for learning chain, in addition to adding random slopes for the generation factor. By doing so, the mixed-effect model will be accounting for variation across chains, in addition to modelling the variation in changes over the generation factor. Taking such an approach to analysis once again highlights the importance, and the need for, sufficiently powered iterated learning experiments.



## 6.4 Concluding remarks

This thesis provides a substantial body of empirical evidence which highlights the dynamic way language is shaped over the course of development and evolution. The emphasis and core focus was to present a view of language as shaped by the brain, where cognitive factors contribute significantly to the way we acquire language, the way we use language, and how languages adapt and evolve in response to pressures introduced by human cognition. The evidence presented here has strengthened our understanding of a variety of contentious issues in the literature around language acquisition and evolution, and whilst there is undoubtedly still questions to be found, that is what makes studying language an endlessly fruitful endeavour, as Müller (1861) wisely notes “*The study of words may be tedious to the school-boy, as breaking of stones is to the wayside laborer; but to the thoughtful eye of the geologist these stones are full of interest;—he sees miracles on the high-road, and reads chronicles in every ditch*” (p.5).



# References

---

- Akita, K. (2013). Constraints on the semantic extension of onomatopoeia. *Public Journal of Semiotics*, 5(1), 21-37.
- Assaneo, M. F., Nichols, J. I., & Trevisan, M. A. (2011). The anatomy of onomatopoeia. *PloS one*, 6(12), e28317.
- Atkinson, Q. D., Meade, A., Venditti, C., Greenhill, S. J., & Pagel, M. (2008). Languages evolve in punctuational bursts. *Science*, 319(5863), 588-588.
- Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology*, 46, 81–86.
- Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1), 31-56.
- Baayen, R. H. (2008). *A practical introduction to statistics using r*. Cambridge, UK: Cambridge University Press.
- Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the

- structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 14(6), 575-589.
- Balota, D. A., Pilotti, M. & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, 29, 639–647.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3), 255-278.
- Belke, E., Brysbaert, M., Meyer, A. S., & Ghyselinck, M. (2005). Age of acquisition effects in picture naming: evidence for a lexical-semantic competition hypothesis. *Cognition*, 96(2), 45-54.
- Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, 113(2), 1001-1024.
- Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 290-311.
- Bickerton, D. (1990). *Language and species*. University of Chicago Press.
- Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 201605782.

- Boersma, P. & Weenink, D. (2009). *Praat: doing phonetics by computer*, Version 6.0.07 [Computer program]. Retrieved December 12, 2014, from <http://www.praat.org>
- Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago Press.
- Boyland, J. T. (1996). *Morphosyntactic change in progress: A psycholinguistic approach*. Unpublished doctoral dissertation, University of California, Berkeley.
- Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). “bouba” and “kiki” in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to westerners. *Cognition*, 126(2), 165–172.
- Brysbaert, M., & Ellis, A. W. (2015). Aphasia and age of acquisition: are early-learned words more resilient?. *Aphasiology*, 1-24.
- Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition*, 13(7-8), 992-1011.
- Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, 104(2), 215-226.
- Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam, the Netherlands: John Benjamins.

- Bybee, J. (2001). *Phonology and language use*. Cambridge, MA: Cambridge University Press.
- Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14, 261-290.
- Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 711-733.
- Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of don't in English. *Linguistics*, 37, 575–596.
- Bybee, J., & Thompson, S. (1997). Three frequency effects in syntax. *Berkeley Linguistics Society*, 23, 378–388.
- Calude, A. S., & Pagel, M. (2011). How do we use language? Shared patterns in the frequency of word use across 17 world languages. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1567), 1101-1107.
- Calude, A. S., & Verkerk, A. (2016). The typology and diachrony of higher numerals in Indo-European: a phylogenetic comparative study. *Journal of Language Evolution*, 1(2), 91-108.
- Campisi, E., & Özyürek, A. (2013). Iconicity as a communicative strategy: Recipient design in multimodal demonstrations for adults and children. *Journal of Pragmatics*, 47, 14–27.

- Carling, G., & Johansson, N. (2014). Motivated language change: processes involved in the growth and conventionalization of onomatopoeia and sound symbolism. *Acta Linguistica Hafniensia*, 46(2), 199-217.
- Cassidy, K. W., & Kelly, M. H. (2001). Children's use of phonology to infer grammatical class in vocabulary learning. *Psychonomic bulletin & review*, 8(3), 519-523.
- Catling, J. C., & Johnston, R. A. (2009). The varying effects of age of acquisition. *The Quarterly Journal of Experimental Psychology*, 62(1), 50-62.
- Catling, J., Dent, K., Preece, E., & Johnston, R. (2013). Age-of-acquisition effects in novel picture naming: A laboratory analogue. *The Quarterly Journal of Experimental Psychology*, 66(9), 1756-1763.
- Chen, C., Gershkoff-Stowe, L., Wu, C., Cheung, H., & Yu, C. (2016). Tracking multiple statistics: Simultaneous learning of object names and categories in English and Mandarin Speakers. *Cognitive Science*.
- Childs, G. T. (1994). African ideophones. In L. Hinton, J. Nichols & J. J. Ohala (Eds.) *Sound Symbolism* (pp. 178–206). Cambridge: Cambridge University Press.
- Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(05), 489-509.
- Christiansen, M. H., & Chater, N. (2015). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 1-

- Christiansen, M. H., & Monaghan, P. (2016), Division of Labor in Vocabulary Structure: Insights From Corpus Analyses. *Topics in Cognitive Science*, 8, 610–624.
- Clark, H. H. (1996). *Using language*. Cambridge University Press: Cambridge.
- Comrie, B. (1981). *Language universals and language typology*. Syntax and Morphology. Oxford: Blackwell.
- Croft, W. (2000). *Explaining language change: An evolutionary approach*. Pearson Education.
- Cuskley, C. (2013). Shared cross-modal associations and the emergence of the lexicon. Doctoral Dissertation, University of Edinburgh.  
<https://www.era.lib.ed.ac.uk/handle/1842/7702>. Accessed 18th February, 2017.
- Cuskley, C., Kirby, S., & Simner, J. (2015). Cross-modality: Reviving iconicity in the evolution of language. In *The evolution of language: Proceedings of the 8th international conference*. London, UK: World scientific.
- Cysouw, M. & Good, J. (2013). Languoid, doculect, and glossonym: Formalizing the notion ‘language’. *Language Documentation & Conservation*, 7, 331–359.
- Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. London: Murray



- Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2016). Wordform Similarity Increases With Semantic Similarity: An Analysis of 100 Languages. *Cognitive Science*.
- De Saussure, F. (1916). *Course in general linguistics*. New York: Columbia University Press.
- Deacon, T. (1997). *The symbolic species*. London: Penguin.
- Dessalles, J. L. (2008). Why is language well designed for communication? [Peer commentary on “Language as shaped by the brain” by M. H. Christiansen & N. Chater]. *Behavioral and Brain Sciences*, 31(05), 518-519.
- Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, 25(2), 108-127.
- Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass*, 6(10), 654-672.
- Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19(10), 603–615.
- Dingemanse, M., Schuerman, W., Reinisch, E., Tufvesson, S., & Mitterer, H. (2016). What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language*, 92(2), e117-e133.
- Eco, U. (1995). *The search for the perfect language*. London, England: Blackwell.

- Edmiston P., Perlman M. & Lupyan G. (2016). The Fidelity Of Iterated Vocal Imitation. In S.G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér & T. Verhoef (eds.) *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*. Available online: <http://evolang.org/neworleans/papers/189.html>
- Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *Journal of Experimental Psychology: Learning, memory, and cognition*, 26(5), 1103.
- Ellis, A.W. (2012). The acquisition, retention, and loss of vocabulary in aphasia, dementia, and other neuropsychological conditions. In M. Faust (Ed.), *The handbook of the neuropsychology of language* (pp. 637-660) Oxford: Blackwell Publishing Ltd.
- Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, 18, 91–126.
- Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, 24(2), 143-188.
- Elsen, H. (1991). *Erstspracherwerb. Der Erwerb des deutschen Lautsystems*. Wiesbaden: DUV.
- Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(05), 429-448.

- Farmer, T.A., Christiansen, M.H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, 103, 12203-12208.
- Fernald, A. (1991). Prosody in speech to children: Prelinguistic and linguistic functions. *Annals of Child Development*, 8, 43–80.
- Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40(4), 296-340.
- Ferrer i Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3), 788-791.
- Fitneva, S. A., Christiansen, M. H., & Monaghan, P. (2009). From sound to syntax: Phonological constraints on children's lexical categorization of new words. *Journal of Child Language*, 36(05), 967-997.
- Ford, M., Bresnan, J. W., & Kaplan, R. M. (1982). A competence-based theory of syntactic closure. In J. W. Bresnan (Ed.) *The mental representation of grammatical relations* (pp. 727-796). Cambridge, MA: MIT Press.
- Fort, M., Martin, A., & Peperkamp, S. (2015). Consonants are more important than vowels in the bouba-kiki effect. *Language and Speech*, 58(2), 247–266.
- Francis, W., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

- Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th annual meeting of the cognitive science society* (pp. 933-938). Washington, DC: Cognitive Science Society.
- Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 80, 748-775.
- Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: where might graphical symbol systems come from?. *Cognitive Science*, 31(6), 961-987.
- Gasser, M. (2004). The origins of arbitrariness in language. In *Proceedings of the Annual Conference of the Cognitive Science Society* (p. 434 - 439). Lawrence Erlbaum.
- Gasser, M., Sethuraman, N., & Hockema, S. (2010). Iconicity in expressives: An empirical investigation. In S. Rice & Newman, J. (Eds.) *Experimental and empirical methods*. Stanford, CA: CSLI Publications.
- Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965), 435-439.
- Gray, R. D., & Jordan, F. M. (2000). Language trees support the express-train sequence of Austronesian expansion. *Nature*, 405(6790), 1052-1055.
- Greenberg, J. (1963). *Universals of language*. Cambridge, MA: MIT Press.

- Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31(3), 441-480.
- Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1509), 3503-3514.
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335-346.
- Hay, J. B., Pierrehumbert, J. B., Walker, A. J., & LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition*, 139, 83-91.
- Hockett, C. (1960). The origin of speech. *Scientific American*, 203, 88-96.
- Hodgson, C., & Ellis, A. W. (1998). Last in, first to go: Age of acquisition and naming in the elderly. *Brain and Language*, 64(1), 146-163.
- Hooper, J. (1976). Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie (Ed.), *Current progress in historical linguistics* (pp. 96–105). Amsterdam: North Holland.
- Horst, J. S., Samuelson, L. K., Kucker, S. C., & McMurray, B. (2011). What's new? Children prefer novelty in referent selection. *Cognition*, 118, 234-244.
- Hruschka, D. J., Christiansen, M. H., Blythe, R. A., Croft, W., Heggarty, P.,

- Mufwene, S. S., Pierrehumbert, J. B., & Poplack, S. (2009). Building social cognitive models of language change. *Trends in cognitive sciences*, 13(11), 464-469.
- Hupp, J. M., & Jungers, M. K. (2013). Beyond words: Comprehension and production of pragmatic prosody in adults and children. *Journal of experimental child psychology*, 115(3), 536-551.
- Hurford, J. R. (2011). *The origins of grammar: Language in the light of evolution II* (Vol. 2). Oxford University Press.
- Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369, 20130298.
- Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, 109(1), 54–65.
- Izura, C., Pérez, M. A., Agallou, E., Wright, V. C., Marín, J., Stadthagen-González, H., & Ellis, A. W. (2011). Age/order of acquisition effects and the cumulative learning of foreign words: A word training study. *Journal of Memory and Language*, 64(1), 32-58.
- Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1), 23-62.

- Jaeger, T. F., & Levy, R. P. (2006). Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems* (pp. 849-856).
- Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition, 13*(7-8), 789-845.
- Joseph, H. S., Wonnacott, E., Forbes, P., & Nation, K. (2014). Becoming a written word: Eye movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition, 133*(1), 238-248.
- Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin, 131*(5), 684.
- Juliano, C., & Tanenhaus, M. K. (1993). Contingent frequency effects in syntactic ambiguity resolution. In *Proceedings of the 15th annual conference of the Cognitive Science Society* (pp. 593-598). Erlbaum Hillsdale, NJ.
- Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*, 137-194.
- Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language, 45*, 229-254.

- Kantartzis, K., Imai, M., & Kita, S. (2011). Japanese sound-symbolism facilitates word learning in English-speaking children. *Cognitive Science*, 35(3), 575–586.
- Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, 99, 349–364.
- Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. Oxford University Press: Oxford.
- Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2), 102-110.
- Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12), 5241-5245.
- Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28, 108-114.
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87-102.
- Köhler, W. (1929). *Gestalt psychology*. New York, NY: Liveright.



- Kovic, V., Plunkett, K., & Westermann, G. (2010). The shape of words in the brain. *Cognition*, 114(1), 19–28.
- Krug, M. (2003). Frequency as a determinant of grammatical variation and change. In G. Rohdenburg, & B. Mondorf (Eds.), *Determinants of grammatical variation in English* (pp. 7–67). Berlin: Mouton de Gruyter.
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978-990.
- Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language*, 83, 152-178.
- Labov, W. (2001). *Principles of Linguistic Change: Social Factors*. Blackwell, Oxford, UK.
- Laing, C. E. (2014). A phonological analysis of onomatopoeia in early word production. *First Language*, 34(5), 387-405.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, 4, 863.
- Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites (Vol. 1)*. Stanford university press.
- Lewis, M. B. (2006). Chasing psycholinguistic effects: A cautionary tale. *Visual*

*Cognition*, 13(7-8), 1012-1026.

Lewis, M. L., & Frank, M. C. (2015). Conceptual complexity and the evolution of the lexicon. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 1138-343).

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, 153, 182-195.

Lewis, M. L., Sugarman, E., & Frank, M. C. (2014). The structure of the lexicon reflects principles of communication. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 845-850).

Lewis, M. P, Simons, G. F., & Fennig, C. D. (eds.). 2016. *Ethnologue: Languages of the World, Nineteenth edition*. Dallas, Texas: SIL International. Online version: <http://www.ethnologue.com>.

Lieberman, E., Michel, J. B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, 449(7163), 713-716.

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in psychology*, 6, 1246.

Lupyan, G., & Casasanto, D. (2015). Meaningless words promote meaningful categorization. *Language and Cognition*, 7(02), 167-193.

Mace, R., & Holden, C. J. (2005). A phylogenetic approach to cultural evolution.

*Trends in Ecology & Evolution*, 20(3), 116-121.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013).

Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2), 313-318.

Marful, A., Gómez-Ariza, C. J., Barbón, A., & Bajo, T. (2016). Forgetting

“Novel” but Not “Dragon”: The Role of Age of Acquisition on

Intentional and Incidental Forgetting. *PloS One*, 11(5), e0155110.

Massaro, D. W., & Perlman, M. (2017). Quantifying Iconicity’s Contribution

during Language Acquisition: Implications for Vocabulary Learning.

*Frontiers in Communication*, 2, 4.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas:

Sound–shape correspondences in toddlers and adults. *Developmental*

*Science*, 9(3), 316–322.

McCormick, K., Kim, J. Y., List, S., & Nygaard, L. C. (2015). Sound to Meaning

Mappings in the Bouba-Kiki Effect. In *Proceedings of the 37th Annual*

*Conference of the Cognitive Science Society* (pp. 1565–1570).

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*.

University of Chicago press.

Monaghan, P. (2014). Age of acquisition predicts rate of lexical evolution.

*Cognition*, 133(3), 530-534.

Monaghan, P. (2017), Canalization of Language Structure From Environmental

- Constraints: A Computational Model of Word Learning From Multiple Cues. *Topics in Cognitive Science*, 9, 21–34.
- Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, 96(2), 143-182.
- Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive psychology*, 55(4), 259-305.
- Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, 140(3), 325–347.
- Monaghan, P., Christiansen, M. H., Farmer, T. A., & Fitneva, S. A. (2010). Measures of phonological typicality: Robust coherence and psychological validity. *The Mental Lexicon*, 5(3), 281-299.
- Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in language learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(5), 1152-1164.
- Monaghan, P., Mattock, K., Davies, R. A., & Smith, A. C. (2015). Gavagai Is as Gavagai Does: Learning Nouns and Verbs From Cross-Situational Statistics. *Cognitive science*, 39(5), 1099-1112.

- Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 20130299.
- Müller, F. M. (1861). *Lectures on the science of language: Delivered at the Royal Institution of Great Britain in April, May, & June 1861*. Project Gutenberg Edition [Ebook 32856].
- Navarrete, E., Pastore, M., Valentini, R., & Peressotti, F. (2015). First learned words are not forgotten: Age-of-acquisition effects in the tip-of-the-tongue experience. *Memory & Cognition*, 43(7), 1085-1103.
- Nettle, D. (1999). Is the rate of linguistic change constant?. *Lingua*, 108(2), 119-136.
- Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition*, 4, 115–125.
- Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009a). Sound to meaning correspondences facilitate word learning. *Cognition*, 112(1), 181–186.
- Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009b). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive Science*, 33(1), 127-146.
- Oliphant, M. (2002). Learned systems of arbitrary reference: The foundation of human linguistic uniqueness. In E. Briscoe (Ed.), *Linguistic evolution*

- through language acquisition: Formal and computational models* (pp. 23–52). Cambridge: Cambridge University Press.
- Onnis, L., & Christiansen, M. H. (2008). Lexical categories at the edge of the word. *Cognitive Science*, 32(1), 184-221.
- Otis, K., & Sagi, E. (2008). Phonoaesthemes: A corpora-based analysis. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 65–70). Austin, TX: Cognitive Science Society.
- Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, 114(2), 173–186.
- Pagel, M. (2009). Human language as a culturally transmitted replicator. *Nature Reviews Genetics*, 10(6), 405-415.
- Pagel, M., & Meade, A. (2006). Estimating rates of lexical replacement on phylogenetic trees of languages. In Forster, P. and Renfrew, C. (Eds.), *Phylogenetic methods and the prehistory of languages* (pp. 173-182). McDonald institute Monographs.
- Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449(7163), 717-720.
- Pagel, M., Atkinson, Q. D., Calude, A. S., & Meade, A. (2013). Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the*

*National Academy of Sciences*, 110(21), 8471-8476.

Peña, M., Mehler, J., & Nespor, M. (2011). The role of audiovisual processing in early conceptual development. *Psychological science*, 22(11), 1419-1421.

Perlman, M., Clark, N., & Johansson Falck, M. (2015). Iconic prosody in story reading. *Cognitive science*, 39(6), 1348-1368.

Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society open science*, 2(8), 150152.

Perniss, P., Thompson, R., & Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in psychology*, 1, 227.

Perry L., Perlman M., Lupyan G., Winter B. & Massaro D. (2016). Early Learned Words Are More Iconic. In S.G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér & T. Verhoef (eds.) *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*. Available online: <http://evolang.org/neworleans/papers/34.html>

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PloS one*, 10(9), e0137147.

Philps, D. (2011). Reconsidering phonæstheses: submorphemic invariance in English 'sn-words'. *Lingua*, 121(6), 1121-1137.

- Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5), 1112-1130.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526-3529.
- Plato. (1971). Cratylus. In Hamilton, E. & Cairns, H. (Eds.), *The collected dialogues of Plato, including the letters*. Vol. 71. Princeton University Press.
- Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Version 3.3.1 Vienna, Austria. URL <https://www.R-project.org>
- Rafferty, A. N., Griffiths, T. L., & Ettlinger, M. (2013). Greater learnability is not sufficient to produce cultural universals. *Cognition*, 129(1), 70-87.
- Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia – a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12), 3–34.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926-1928.
- Scheibman, J. (2000). I dunno: A usage-based account of the phonological



- reduction of don't in American English conversation. *Journal of Pragmatics*, 32(1), 105-124.
- Schmidtke, D., Conrad, M., & Jacobs, A. M. (2014). Phonological iconicity. *Frontiers in psychology*, 5, 80.
- Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, 113(2), 226-233.
- Senghas, A., Kita, S., & Özyürek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, 305(5691), 1779-1782.
- Shapiro, B. J. (1969). The subjective estimate of relative word frequency. *Journal of Verbal Learning and Verbal Behavior*, 8, 248–251.
- Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72(2), B11-B21.
- Shintel, H., Nusbaum, H. C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language*, 55, 167–177.
- Simner, J., Cuskley, C., & Kirby, S. (2010). What sound does that taste? Cross-modal mappings across gustation and audition. *Perception*, 39(4), 553-569.
- Smith, C. A. (2017). Tracking semantic change in fl-monomorphemes in the Oxford English Dictionary. *Journal of Historical Linguistics*, 6(2), 165-

200.

- Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558–1568.
- Spector, F., & Maurer, D. (2009). Synesthesia: a new approach to understanding the development of perception. *Developmental psychology*, 45(1), 175.
- Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73, 971–995.
- St Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, 33(7), 1317-1329.
- Steyvers, M., & Tenenbaum, J. B. (2005). The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1), 41-78.
- Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*, 96(4), 452-463.
- Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. In Haspelmath, M., & Tadmor, U. (Eds.), *Loanwords in the world's languages: A comparative handbook* (pp.55-75). Berlin: De Gruyter Mouton.
- Tamariz, M. (2004). *Exploring the adaptive structure of the mental lexicon*.

- Unpublished doctoral dissertation, University of Edinburgh.
- Tamariz, M. (2008). Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, 3(2), 259-278.
- Tamariz, M. (2017). Experimental Studies on the Cultural Evolution of Language. *Annual Review of Linguistics*, (0).
- Tamminen, J., & Gaskell, M. G. (2008). Newly learned spoken words show long-term lexical competition effects. *The Quarterly Journal of Experimental Psychology*, 61(3), 361-371.
- Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology*, 44(4), 929.
- Thomason, S. G., & Kaufman, T. (1992). *Language contact, creolization, and genetic linguistics*. Berkley: University of California Press.
- Thompson, R. L., Vinson, D. P., Woll, B., & Vigliocco, G. (2012). The road to language learning is iconic evidence from british sign language. *Psychological science*, 23(12), 1443-1448.
- Thompson, R. L., Vinson, D. P., Woll, B., & Vigliocco, G. (2012). The road to language learning is iconic evidence from british sign language. *Psychological science*, 23(12), 1443-1448.
- Urban, M. (2011). Asymmetries in overt marking and directionality in semantic change. *Journal of Historical Linguistics*, 1, 3-47.

- van der Loo M (2014). The stringdist package for approximate string matching. *The R Journal*, 6, 111-122. <https://CRAN.R-project.org/package=stringdist>
- Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67(6), 1176-1190.
- Van Loon-Vervoorn, W. A. (1989). *Eigenschappen van basiswoorden: (Properties of basic words): proefschrift*. Swets & Zeitlinger.
- Vejdemo, S., & Hörberg, T. (2016). Semantic Factors Predict the Rate of Lexical Replacement of Content Words. *PloS one*, 11(1), e0147924.
- Vinson, D. P., Cormier, K., Denmark, T., Schembri, A., & Vigliocco, G. (2008). The British Sign Language (BSL) norms for age of acquisition, familiarity, and iconicity. *Behavior Research Methods*, 40(4), 1079-1087.
- Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, 21(1), 21–25.
- Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5), 2020.

- Wilkins, J. (1668). *An essay towards a real character and a philosophical language*. London, England: Gellibrand.
- Winter, B., Thompson, G., & Urban, M. (2013). Cognitive factors motivating the evolution of word meanings: Evidence from corpora, behavioral data and encyclopedic network structure. In: E. A. Cartmill, S. Roberts, H. Lyn, & H. Cornish (Eds.), *10th International Conference on the Evolution of Language* (pp. 353-360). New Jersey: World Scientific.
- Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution*, 1(1), 7-18.
- Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414–420.
- Zevin, J. D., & Seidenberg, M. S. (2004). Age-of-acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory & Cognition*, 32(1), 31-38.
- Zipf, G. (1936). *The psychobiology of language*. Routledge, London.
- Zipf, G. (1949). *Human behavior and the principle of least effort*. Addison-Wesley, New York.



# Appendix

---

## A1. Tables from Experiment 1

**Table A1.1.** Main model selection. The table provides Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and log-likelihood (logLik) for several potential models fit to the data for Experiment 1. For all models, the `glmer()` call was `Response ~ [Fixed effects] + (1|Subject) + (1|Sound)`, and fit a binomial model (i.e., all models used the same outcome variable and random effects).

Model	Fixed effects	AIC	BIC	LogLik	$\chi^2$	<i>p</i>	Preferred model
1	-	18201	18223	-9097.4	-	-	-
2	1 + condition	18204	18241	-9096.9	0.9655	0.6171	1
3	1 + block	18051	18081	-9021.3	152.14	<0.0001	3
4	3 + congruency	17995	18033	-8992.5	57.633	<0.0001	4
5	4 + same or different shape condition	17996	18042	-8992.2	0.4949	0.4817	4
6	4 + condition x congruency	17994	18062	-8988.1	8.7971	0.0664	4
7	4 + condition x same or different shape condition	18002	18078	-8991.2	2.5736	0.7654	4
8	4 + congruency x same or different shape condition	17962	18015	-8974.1	36.753	<0.0001	8
9	8 + condition x congruency x same or different shape condition	17947	18060	-8958.3	31.511	<0.001	9

**Table A1.2.** Summary of the Generalized Linear Mixed-effects Model of (log odds) accuracy of response over blocks, experimental conditions, congruency and same or different shape condition. R syntax for final model is: `glmer(accuracy ~ block + condition + congruency + learning_type + condition*congruency*learning_type + (1 | Subject) + (1 | Sound)`

Fixed effects	Estimated coefficient	SE	Wald confidence intervals		z	Pr(> z )
			2.50%	97.50%		
(Intercept)	0.2388	0.0720	0.0978	0.3798	3.3180	0.0009
Block effect	0.1983	0.0161	0.1667	0.2298	12.3280	<0.0001
Congruency (congruent vs. incongruent)	-0.4736	0.0544	-0.5802	-0.3671	-8.7120	<0.0001
Same or different shape condition (categorical vs. individual)	-0.2088	0.0536	-0.3139	-0.1038	-3.8980	<0.0001
Experimental condition (linear)	-0.1619	0.0973	-0.3526	0.0289	-1.6630	0.0963
Experimental condition (quadratic)	-0.1521	0.0964	-0.3410	0.0368	-1.5780	0.1145
Congruency:same or different shape condition interaction	0.3694	0.0746	0.2232	0.5156	4.9530	<0.0001
Experimental condition (linear):congruency interaction	0.1672	0.0936	-0.0162	0.3506	1.7870	0.0740
Experimental condition (quadratic):congruency interaction	0.4260	0.0902	0.2492	0.6027	4.7230	<0.0001
Experimental condition (linear):same or different shape condition interaction	0.2543	0.0942	0.0696	0.4390	2.6990	0.0070
Experimental condition (quadratic):same or different shape condition interaction	0.2384	0.0912	0.0597	0.4170	2.6150	0.0089
Experimental condition (linear):congruency:same or different shape condition interaction	-0.3918	0.1316	-0.6497	-0.1340	-2.9780	0.0029
Experimental condition (quadratic):congruency:same or different shape condition interaction	-0.5171	0.1266	-0.7652	-0.2689	-4.0840	<0.0001
Random effects						
Groups	Name	Variance	Std.Dev.			
Subject effect on intercepts	(Intercept)	0.12	0.35			
Item effect (objects) on intercepts	(Intercept)	0.01	0.09			
	AIC	BIC	logLik	deviance		
	17946.7	18059.7	-8958.3	17916.7		

13824 observations, 72 participants, 16 sound stimuli



## A2. Across experiment comparisons (Experiment 1 to Experiments 2 & 3)

To distinguish between the three different experiments, the following terms will be used to refer to the individual experimental conditions: Experiment 1 – *Fully congruent*, experiment 2 – *no relationship*, [Chapter 2](#) – mixed (with distinctions made between congruent mappings – *mixed (congruent)* and incongruent mappings – *mixed (incongruent)*).

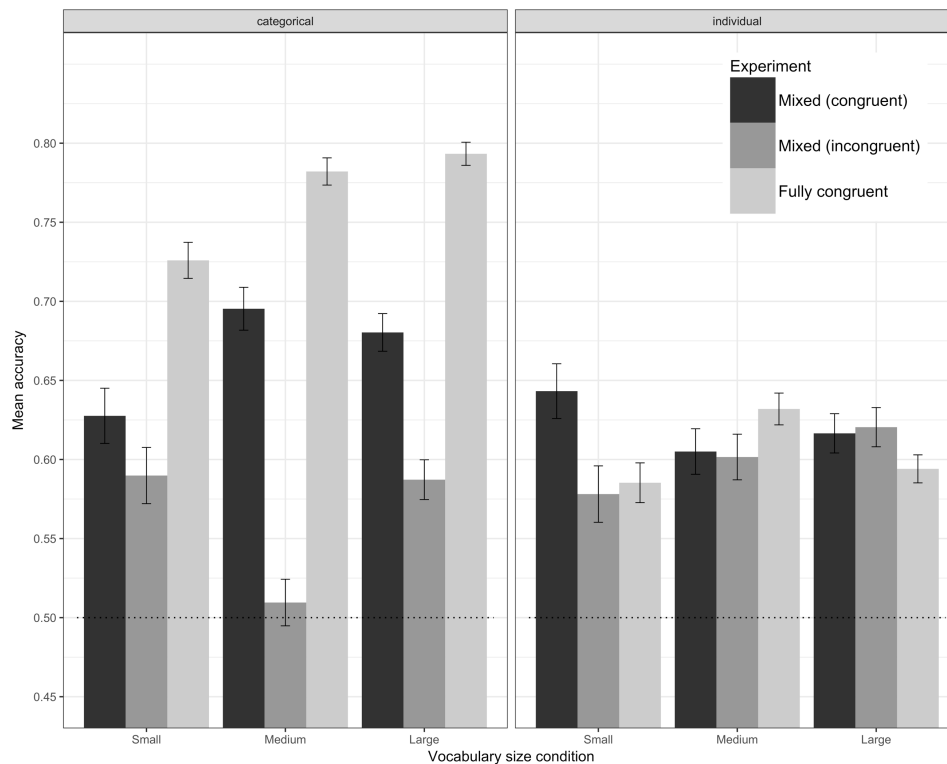
### Mixed (congruent/incongruent) and fully congruent

The addition of experiment as a fixed effect revealed a significant improvement to model fit ( $\chi^2(2) = 80.442, p < .001$ ), with accuracy in the fully congruent language being significantly greater than the mixed (congruent) condition (estimate = -.21,  $SE = .08, z = -2.71, p = .007$ ) and mixed (incongruent) condition (estimate = -.48,  $SE = .08, z = -6.36, p < .001$ ). Furthermore, the addition of the three-way interaction between vocabulary size x presentation type x experimental condition also significantly improved model fit ( $\chi^2(12) = 50.22, p < .001$ ).

To understand this interaction, we then carried out further analyses on the separate presentation type trials (categorical or individual learning) and the different vocabulary sizes. For the categorical trials, the inclusion of experimental condition significantly improved model fit for all the vocabulary size conditions (small:  $\chi^2(2) = 10.034, p = .007$ ; medium:  $\chi^2(2) = 113.12, p < .001$ ; large:  $\chi^2(2) =$

49.323,  $p < .001$ ). The analyses performed on the small vocabulary size data demonstrated that accuracy in the fully congruent language was significantly greater than both the mixed congruent and incongruent conditions (estimate =  $-.55$ ,  $SE = .23$ ,  $z = -2.37$ ,  $p = .018$  and estimate =  $-.73$ ,  $SE = .23$ ,  $z = -3.12$ ,  $p = .002$  respectively). This was also the case for the medium vocabulary size (mixed (congruent): estimate =  $-.52$ ,  $SE = .18$ ,  $z = -2.95$ ,  $p = .003$ ; mixed (incongruent): estimate =  $-1.36$ ,  $SE = .18$ ,  $z = -7.69$ ,  $p < .001$ ) and the large vocabulary size (mixed (congruent): estimate =  $-.72$ ,  $SE = .20$ ,  $z = -3.55$ ,  $p < .001$ ; mixed (incongruent): estimate =  $-1.15$ ,  $SE = .20$ ,  $z = -5.68$ ,  $p < .001$ ). See [Figure A2.1](#) for results.

For the individual learning trials, the inclusion of experimental condition significantly improved model fit in the small vocabulary size ( $\chi^2(2) = 7.87$ ,  $p = .020$ ), however there was no significant difference in accuracy between the fully congruent and mixed incongruent conditions (estimate =  $-.03$   $SE = .13$ ,  $z = -.24$ ,  $p = .813$ ) and a marginally significant difference between the fully congruent and mixed congruent conditions (estimate =  $.25$   $SE = .13$ ,  $z = 1.93$ ,  $p = .054$ ), with the only significant difference found being between mixed congruent and mixed incongruent conditions (as reported in [Chapter 2](#)). This indicates that accuracy was only significantly different when the language was mixed. For the medium and large vocabulary sizes, the addition of experimental condition did not significantly improve model fit (all  $p$ 's  $> .05$ ), indicating that



**Figure A2.1.** Mean accuracy of responses for mixed (congruent/incongruent) and fully congruent experimental conditions. Error bars show SEM. Dotted line shows 50% chance level.

there was no difference between any of the experimental conditions during individual learning trials as vocabulary size increased. See [Figure A2.1](#) for results.

These results demonstrate that a fully sound-symbolic language provides the learner with the greatest benefit when learning to distinguish across categorical distinctions, in this case rounded/angular shapes, a pattern of results that is observed regardless of vocabulary size. This is in line with our hypothesis, that sound-symbolism provides information about the category of the meaning being referred to and that this information can be used by the learner to effectively learn the language. However, there is only a benefit to learning

individual form-meaning mappings when the vocabulary size is small and the language comprises both congruent and incongruent mappings, with the advantage being observed for the congruent mappings. As the vocabulary size increases there are no observable differences between a fully congruent or a mixed language. Although our hypothesis was that sound-symbolism should benefit individual word learning in a small vocabulary size, learning from a fully congruent language only provided intermediate performance between the mixed congruent and incongruent condition. This may suggest that having a balance between sound-symbolic and non-sound-symbolic mappings could be important for the learning of individuated form-meaning mappings, instead of an exclusively sound-symbolic language.

### **Mixed (congruent/incongruent) and no relationship**

The addition of experiment as a fixed effect revealed a significant improvement to model fit ( $\chi^2(2) = 61.1, p < .001$ ), with accuracy in the mixed congruent language being significantly greater than the no relationship condition (estimate = .22,  $SE = .06, z = 3.40, p < .001$ ). There was no significant difference between the no relationship and mixed incongruent conditions (estimate = -.06,  $SE = .06, z = -.89, p = .376$ ). Furthermore, the addition of the three-way interaction between vocabulary size x presentation type x experimental condition also significantly improved model fit ( $\chi^2(12) = 35.38, p < .001$ ).

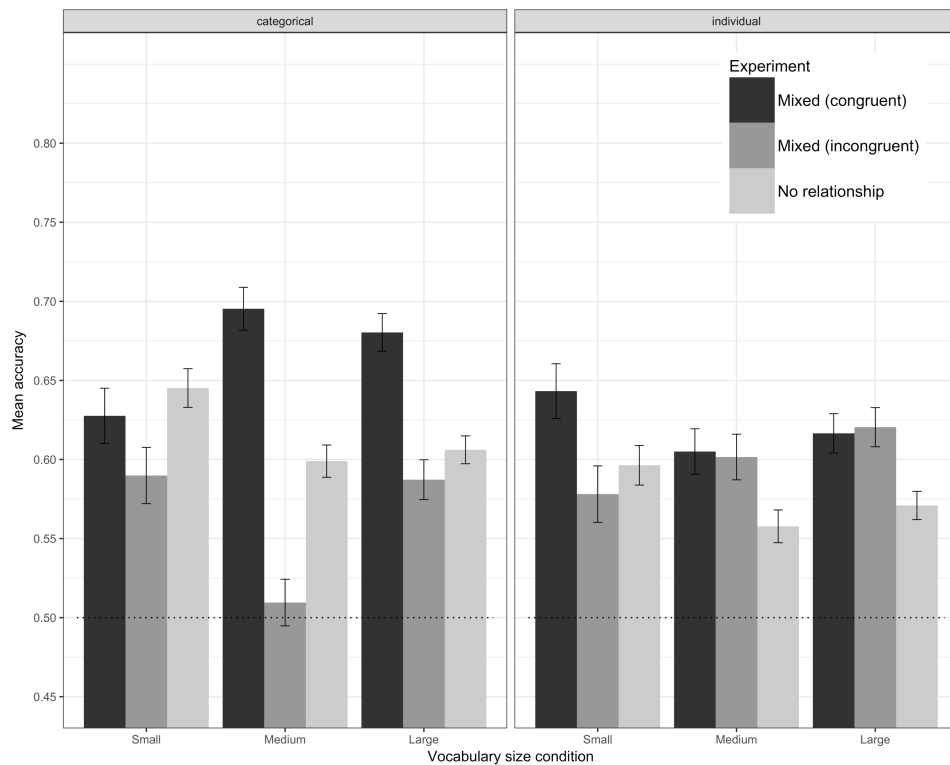
To understand this interaction, we then carried out further analyses on the separate presentation type trials (categorical or individual learning) and the

different vocabulary sizes. For the categorical trials, the inclusion of experimental condition did not significantly improve model fit for the small vocabulary size condition ( $\chi^2(2) = 3.36, p = .186$ ), indicating that there were no differences in accuracy between the three experimental conditions. However, for the medium and large vocabulary sizes there was a significant improvement to model fit (medium:  $\chi^2(2) = 86.78, p < .001$ ; large:  $\chi^2(2) = 31.17, p < .001$ ), with accuracy in the mixed congruent condition greater than the no relationship condition (medium; estimate = .44,  $SE = .13, z = 3.42, p < .001$ ; large: estimate = .36,  $SE = .14, z = 2.55, p = .011$ ). There was however significantly higher accuracy in the no relationship condition, when compared to the mixed incongruent condition (estimate = .38,  $SE = .13, z = -2.96, p = .003$ ), although this effect was not found in the large vocabulary size ( $p > .05$ ). See [Figure A2.2](#) for results.

For the individual learning trials, the inclusion of experimental condition significantly improved model fit in the small vocabulary size ( $\chi^2(2) = 7.25, p = .027$ ), however there was no significant differences between the no relationship condition and either of the mixed congruent or mixed incongruent conditions (both  $p$ 's  $> .05$ ), with the significant effect only being found when comparing mixed congruent to mixed incongruent (as reported in [Chapter 2](#)). Furthermore, the addition of experimental condition did not significantly improve model fit for the medium or large vocabulary sizes (both  $p$ 's  $> .05$ )<sup>13</sup>, indicating that there was no difference in accuracy between the mixed congruent, mixed incongruent or no

---

<sup>13</sup> Note that for the large vocabulary size there was a marginal effect when experimental condition was added ( $\chi^2(2) = 5.41, p = .067$ ).



**Figure A2.2.** Mean accuracy of responses for mixed (congruent/incongruent) and no relationship experimental conditions. Error bars show SEM. Dotted line shows 50% chance level.

relationship conditions. These results indicate that there was no difference in accuracy between any of the experimental conditions during individual learning trials, apart from the difference between mixed congruent and mixed incongruent. See [Figure A2.2](#) for results.

These results demonstrate that when sound-symbolism was present in the form-meaning mappings (in the mixed congruent condition), then overall accuracy was greater than the conditions where it was not present (mixed incongruent and no relationship). However, when the vocabulary size is small, there appears to be no benefits for sound-symbolic mappings for learning to distinguish between categories or individuated meanings, when compared to

mappings with no relationship between form and meaning. As the vocabulary size increases, there is a benefit for sound-symbolic mappings over those with no relationship, but this is only observed for categorical learning, not the learning of individual meanings. We therefore did not find any evidence that indicates an advantage for learning of individual meanings as the vocabulary size increases when mappings had no relationship between form and meaning.

